Terakreditasi SINTA Peringkat 2
Surat Keputusan Direktur Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Nomor: 158/E/KPT/2021
masa berlaku mulai Volume 5 Nomor 2 Tahun 2021 sampai Volume 10 Nomor 1 Tahun 2026

Terbit online pada laman web jurnal: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 6 No. 1 (2022) 92 - 99 ISSN Media Elektronik: 2580-0760

Gradient Boosting Machine, *Random Forest* dan *Light GBM* untuk Klasifikasi Kacang Kering

Indrawata Wardhana[1], Musi Ariawijaya[2] , Vandri Ahmad Isnaini[3], Rahmi Putri Wirman[4]
[1,2]Sistem Informasi, Fakultas Sains dan Teknologi, UIN Sulthan Thaha Saifuddin Jambi
[3,4]Fisika, Fakultas Sains dan Teknologi, UIN Sulthan Thaha Saifuddin Jambi
[1]indrawataw@uinjambi.ac.id, [2]musi@uinjambi.ac.id, [3]vandri@uinjambi.ac.id, [4]rahmi@uinjambi.ac.id

Abstract

Bean seed classification is critical in determining the quality of beans. Previously, the same dataset was tested using the MLP, SVM, KNN, and DT algorithms, with SVM producing the best results. The purpose of this study is to determine the most effective model through the use of the BoxCox transformation selection feature and the random forest (RF) algorithm, as well as the gradient boosting machine (GBM), light GBM, and repeated k-folds evaluation model. The bean dataset is available on the UCI Repository website. The BoxCox transformation and repeated k-folds improved the classification prediction's accuracy. The model is used in the optimal training phase for a random forest with decision tree parameters 50 and depth 10, a gradient boosting machine model with a learning rate of 1, and a light gradient boosting machine model with a learning rate of 0.5 and estimator of 500. The best training accuracy results are obtained with light GBM. which is 99 percent accurate, but only 91 percent accurate in terms of validation. According research, the Barbunya, Bombay, Cali, Dermason, Horoz, Seker, and Sira beans classes provided accuracy values of 91 percent, 100 percent, 92 percent, 92 percent, 95 percent, 94 percent, and 84 percent, respectively.

Keywords: GBM, RF, LightGBM, Bean Classification, BoxCox

Abstrak

Klasifikasi biji kacang sangat penting dalam penentuan mutu kacang. Dataset yang sama sebelumnya telah diuji menggunakan algoritma MLP, SVM, KNN dan DT, dimana SVM memberikan hasil yang paling baik. Penelitian ini bertujuan untuk melihat model yang paling efektif dengan menggunakan fitur seleksi transformasi *BoxCox* dan algoritma *Random Forest* (RF), *Gradient Boosting Machine* (GBM), *light* GBM serta model evaluasi *repeated k-folds*. Dataset kacang berasal dari website UCI *Repository*. Didapatkan bahwa transformasi *BoxCox* dan *repeated k-folds* meningkatkan akurasi dari prediksi klasifikasi. Penggunaan pada model pada fase training terbaik untuk *random forest* dengan paramater pohon keputusan 50 dan depth 10, model *gradient boosting machine* pada *learning rate* 1, dan model *light gradient boosting* pada *learning rate* 0,5 dan estimator 500. Light GBM memberikan hasil akurasi training terbaik yakni 99 persen namun akurasi validasi hanya 91 persen. Dari prediksi tersebut, didapatkan hasil bahwa kelas kacang Barbunya, Bombay, Cali, Dermason, Horoz, Seker, dan Sira memberikan nilai akurasi berturut-turut yakni 91%, 100%, 92%, 92%, 95%, 94% dan 84%.

Kata kunci: : GBM, RF, Light GBM, klasifikasi kacang, *BoxCox*

1. Pendahuluan

Penentuan klasifikasi biji-bijian merupakan faktor yang penting sekali dalam menentukan mutu biji-bijian dan telah banyak dilakukan dengan berbagai metode oleh para ahli. Beberapa klasifikasi biji menggunakan peralatan laboratorium telah dilakukan yakni menggunakan [1] spektroskopi inframerah, klasifikasi biji kakao dengan analisis kromatografi [2] dan spectrometer inframerah [3]. Untuk meningkatkan hasil analisis klasifikasi, metode penggabungan antara pengukuran laboratorium dan analisa statistik juga dilakukan. Kurniawan dkk menggunakan *Near Infrared* (NIR) Spektroskopi dalam klasifikasi biji kakau dengan lima spektrum data *pretreatment* yakni penghalusan, turunan pertama, turunan kedua, *Multiple Scatter Correction (MSC), Standard Normal Variate (SNV),* dan *Principle Component Analysis (PCA)* [4] . Klasifikasi biji kakau menggunakan kromatografi analisis yang kemudian diolah datanya menggunakan *k-means*,

Diterima Redaksi: 04-12-2021 | Selesai Revisi: 26-01-2022 | Diterbitkan Online: 27-02-2022

principal component analysis, dan *Discriminant Analysis* (DA) dapat mengelompokkan biji tersebut berdasarkan 6 jenis karakteristik kimia [5]. Kemudian penentuan keberlangsungan hidup biji kacang kedelai menggunakan *Fluorescence hyperspectral imaging* (FHSI) dan CARS-SVM-AdaBoost berhasil dideteksi dengan akurasi sebesar 100 persen [6].

Metode analisis serta perhitungan pada *machine learning* dan *image recognition* kacang kering dapat diidentifikasi berdasarkan panjang, bentuk, besar, dan aspek fisik lainnya. Teknik *computer vision* juga dapat digunakan untuk menentukan klasifikasi sampel seperti membedakan jenis bibit kacang kering[7], klasifikasi warna kulit kacang [8], klasifikasi biji kopi [9][10], biji kopi dengan ampas [11] menggunakan klusterisasi *k-means.*

Ada beberapa teknik identifikasi sampel atau objek pada *machine learning*. Transformasi data *BoxCox* banyak digunakan pada *machine learning*. Penggunaan *BoxCox* untuk mengurangi anomali seperti *non-additivity, non-normality,* dan *heteroscedasticity*[12]. Metode *BoxCox* membantu untuk memilih respon tranformasi untuk meyakinkan validitas data dari sebuah distribusi Gaussian [13]. *BoxCox* juga membantu peneliti secara cepat untuk mencari transformasi normalisasi yang optimal untuk tiap variabel[14]. Terdapat dua model seleksi dari transformasi dan aplikasi nya pada model regresi linier [15].

Random Forest (RF) banyak digunakan pada klasifikasi dan regresi, dari penelitian diperoleh hasil bahwa model RF lebih akurat dari akurasi *biomass* gandum dibandingkan dengan *Support Vector Regression* (SVR) dan *Artificial Neural Network* (ANN) pada tiap tahapan, dan kekokohan sama baiknya dengan SVR, namun lebih baik dari ANN. Klasifikasi RF digunakan pada penyakit getah bening, dengan memadukan fitur seleksi algoritma genetik, didapatkan akurasi sebesar 92,2 % [16][17]. Penggunaan RF dalam mengidentifikasi 23 panjang gelombang yang berkaitan dengan struktur tumbuhan dan konten air [18]. Random Forest juga digunakan untuk memprediksi mutasi kanker dari dataset *genomic* [19]. Akurasi RF sangat akurat sebesar 96,57% dan AUC besar dari 98% saat mengenali *host* tropis dari protein influenza individu [20].

Gradient Boosting Machine (GBM) memiliki beberapa keunggulan dari metode *machine learning* lain. Penelitian didapatkan bahwa GBM meningkatkan akurasi prediksi R kuadrat dan RMSE lebih dari 80 persen dibandingkan dengan model terbaik industri yakni algoritma *random forest* dan regresi linier [21]. GBM juga digunakan pada prediksi waktu pergi dan kedatangan, dimana GBM memiliki kelebihan untuk prediksi waktu keberangkatan yang bebas memilih [22]. Beberapa keluarga dari algoritma *gradient boosting* diuji pada kecepatan dan akurasi. Uji dilakukan pada *CatBoost, eXtreme Gradient Boosting* (XGBoost),, *random forests, LightGBM, dan gradient boosting.* Hasil komparasi, diindikasikan bahwa CatBoost merupakan hasil terbaik untuk akurasi dan AUC, walaupun perbedaannya kecil. *Light Gradient Boosting Machine* (LightGBM) tercepat dari semua metode. Dan XGBoost menempati posisi kedua untuk akurasi dan kecepatan training [23].

Light GBM mempercepat waktu proses 20 kali lipat dari fase pelatihan *Gradient Boosting Decision Tree* (GBDT) konvensional dengan akurasi yang sama [24]. Penggunaan LightGBM pada pinjaman jaringan p2p, dengan hasil yang lebih baik dari XGBoost [25]. Pada prediksi miRNA penderita kanker payudara, menggunakan beberapa teknik *machine learning* yakni XGBoost, *Random Forest*, dan lightGBM , diperoleh bahwa LightGBM dari beberapa aspek seperti akurasi dan kecepatan unggul dari dua teknik lainnya [26].

Pada penelitian ini melakukan komparasi terhadap akurasi prediksi pada tiga *algoritma gradient boosting machine, random forest* dan *Light GBM* menggunakan fitur seleksi *BoxCox*. Komparasi ini akan diuji pada klasifikasi dataset kacang kering.

2. Metode Penelitian

Penelitian ini dirancang berdasarkan *flowchart* pada gambar 1, tiap tahapan dari *flowchart* tersebut dijelaskan dalam bentuk sub bagian berikut.

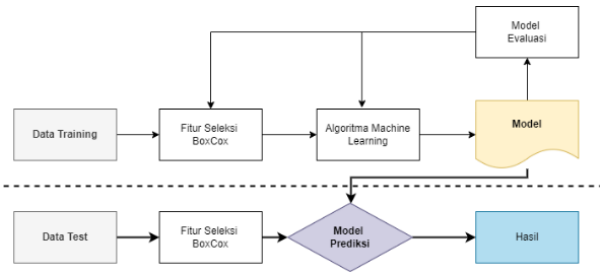
2.1. Dataset Kacang

Data berasal dari website UCI *Repository Dry Bean dataset* [27] [7].Tujuh jenis kacang kering digunakan pada penelitian ini. Fitur yang digunakan seperti bentuk, dimensi, tipe dan struktur. Dataset ini juga di lengkapi dengan gambar dari 7 jenis kacang, namun pada penelitian ini tidak menggunakan dataset gambar tersebut, hanya berdasarkan kelas kacang yang tersedia pada kolom ke 17. Klasifikasi terdiri dari 16 fitur, 12 dimensi, dan 4 bentuk dari permukaan dengan total baris sebanyak 13.611. Beberapa variabel yang digunakan berurutan dalam dataset yakni:

1. Area (A)
2. Perimeter (P)
3. Panjang sumbu utama (L)
4. Panjang sumbu minor (l)
5. Aspek rasio (K)
6. Eksentrisitas (Ec)
7. Area cembung (C)
8. Diameter ekuivalen (Ed)
9. Luas (Mis)
10. Soliditas (S)
11. Kebulatan (R)
12. *Compactness* (CO
13. Faktor Bentuk1 (SF1)
14. Faktor Bentuk2 (SF2)
15. Faktor Bentuk3 (SF3)
16. Faktor Bentuk4 (SF4)

DOI: https://doi.org/10.29207/resti.v5i1.3682

17. Kelas kacang (Barbunya, Bombay, Dermosan, Horoz ,Seker, Cali dan Sira).



Gambar 1. Flowchart Metode Klasifikasi

Korelasi pearson digunakan untuk melihat hubungan antara variabel. Dimana koofisien korelasi Pearson memiliki perumusan berikut ini.

$$Pearson = \frac{cov(x,y)}{stdev(c)\ x\ stdev(y)} \quad (1)$$

Dimana *cov* merupakan covarian, *stdev* adalah standar deviasi. Untuk menginterprestasi koofisien korelasi tersebut dapat dilihat pada tabel berikut :

Tabel 1.Tabel nilai koofisen korelasi dan interprestasi

Korelasi Lansung	Korelasi tidak lansung	Kekuatan Relasi
0.0	0.0	Tidak ada
0.1	-0.1	Lemah/kecil
0.3	-0.3	Menengah
0.5	-0.5	Kuat
1.0	-1.0	Terbaik

2.2. Algoritma Klasifikasi

Proses prediksi pada tes dataset menggunakan *Random forest, Gradient Boosting Machine* dan *Light Gradient Boosting Machine*.

Pada klasifikasi *Random forest* digunakan Gini Index sebagai atribut pengukuran seleksi dimana mengukur ketidakmurnian dari atribut yang berkaitan dengan kelas. Untuk training set T, seleksi satu kasus (piksel) dengan acak dan itu berhubungan dengan beberapa kelas Ci. Gini index dapat di tulis dalam rumusan 2 [28] berikut ini :

$$\sum\sum_{j\neq i}(f(C_i,T)/|T|)(f(C_j,T)/|T|) \quad (2)$$

dimana f (Ci, T)/|T| merupakan probabilitas dari dari seleksi kasus yang berhubungan dengan kelas Ci. Setiap perkembangan dari pohon menuju kedalaman maksimum pada setiap pelatihan data menggunakan kombinasi fitur.

Klasifikasi GBM tujuan nya adalah mencari model *additive* yang meminimalkan fungsi *loss* [21]. Persamaan algortima untuk meningkatkan regresi *trees* dapat digeneralkan pada persamaan 3, dimana model akhir dari model penambahan bertahap sederhana dari nilai b .

$$f(x) = \sum_{b=1}^{B} f^b(x) \quad (3)$$

2.3. Normalisasi data *BoxCox*

BoxCox berfungsi sebagai transformasi pada data tidak normal yang kemudian diubah menjadi bentuk normal. Normalitas merupakan asumsi penting pada berbagai teknik statistik. Transformasi ini dilakukan dengan menggunakan estimasi kemungkinan maksimum untuk mengestimasi eksponen daya lambda, dimana nilainya berkisar antara -5 dan 5. Persamaannya dapat dilihat sebagai berikut.

$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & jika\ \lambda \neq 0 \\ \log y & jika\ \lambda = 0 \end{cases} \quad (4)$$

Dimana lambda merupakan nilai antara -5 dan 5.

2.4. Evaluasi

Tahapan evaluasi terhadap metode klasifikasi dilakukan untuk mengetahui kinerja dari tiap algoritma.
Evaluasi menggunakan *confusion matrix* yang terdiri dari tujuh kelas yakni Barbunya, Bombay, Cali, Dermason, Horoz, Seker, dan Sira. Adapun *confusion matrix* untuk kelas kacang kering tersebut ditunjukkan pada Tabel 2.

Tabel 2.Confusion Matrix Kelas Kacang

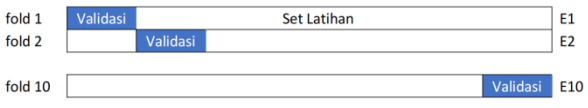
Eksperimen	Kelas Prediksi	
Kelas Aktual	Positif	Negatif
Positif	TP	TN
Negatif	FP	FN

Dimana TN, FN , TP dan FP merupakan kependekan dari *true negative, false negative, true positive* dan *false positive*

Pada tahapan ini, metode yang digunakan adalah *k-fold repeated cross-validation* dalam mengevaluasi performa klasifikasi dataset training dan test. Pengulangan dilakukan sebanyak n kali untuk 10 *k-fold* sesuai persamaan 5 .

$$E = \frac{1}{10}\sum_{i=1}^{10} E_i \quad (5)$$

dimana E merupakan error dari setiap iterasi. Iterasi dari tiap *fold* dapat di gambarkan pada gambar 2.



Gambar 2. 10 k-folds repeated cross-validation

Precision, recall, dan *f1 score* digunakan sebagai metrik dalam evaluasi performa dari klasifikasi. Metrik tersebut diformulasikan pada rumus 6 dan 7 .

$$Precision = \frac{TP}{TP+FP}\ x\ 100\% \quad (6)$$

$$Recall = \frac{TP}{TP+FN}\ x\ 100\% \quad (7)$$

DOI: https://doi.org/10.29207/resti.v5i1.3682

$$f1 = \frac{(1+\beta^2)\ x\ precision\ x\ recall}{(\beta^2\ x\ precision) + recall}\ x\ 100\% \quad (8)$$

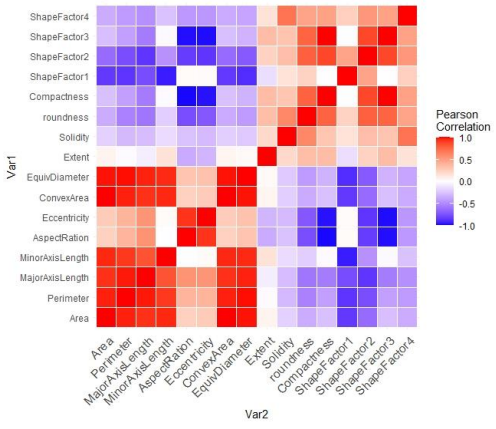
2.5. Peralatan

Penelitian ini menggunakan *hardware* spesifikasi teknis sebagai berikut : Intel VGA HD Graphics 3000, Processor Intel i5, SSHD 1 Tb, dan RAM DDR3 8 Gb. Pengolahan data menggunakan bahasa Python dan R dengan menggunakan *software* Anaconda Jupyter berbasis web.

3. Hasil dan Pembahasan

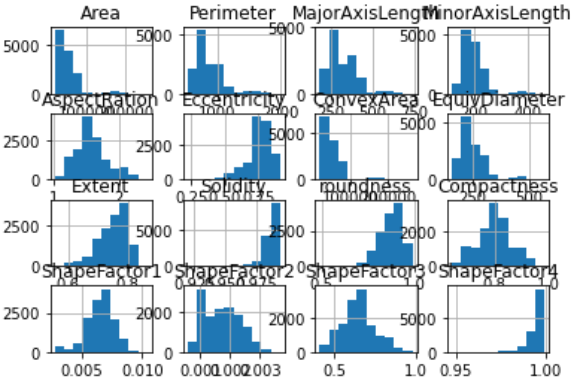
3.1. Korelasi Variabel

Dataset diolah terlebih dahulu menjadi data latih dan data uji. Pembagiannya menjadi tiga bagian yakni: 70:30,80:20, dan 90:10 dengan porsi terbesar pada data latih. Pengecekan nilai kosong dilakukan pada dataset, dimana tidak terdapat data kosong. Hubungan antar variabel pada dataset dianalisis dengan persamaan Pearson. Hubungan antar variabel tersebut dapat dilihat pada Gambar 3. Hubungan terdiri dari relasi langsung dan relasi tidak langsung. Hubungan langsung berada pada kisaran 0 – 1 dan tidak lansung antara 0 dan -1. Beberapa variabel memiliki hubungan langsung di atas 0.5 yakni Area, Perimeter, MajorAxisLength, MinorAxisLength, ConvexArea, EquivDiameter, dan Compacness. Korelasi tidak langsung di bawah -0.5 yakni variabel: Compactness, ShapeFactor 1, ShapeFactor2. Jumlah kelas kacang berdasarkan jenisnya yakni: Barbunya 1322, Bombay 522, Cali 1630, Dermason 3546, Horoz 1928, Seker 2027, dan Sira 2636.



Gambar 3. Korelasi Pearson antar variabel

Dari Gambar 4, didapatkan hanya beberapa variabel saja yang nilainya terdistribusi normal yakni ShapeFactor1, ShapeFactor2, ShapeFactor3, dan Compactness yang memiliki skewness simetris. Sedangkan skewness positif terdapat pada variabel: Area, Perimeter, MinorAxisLength. Berbeda dengan variabel: *Eccentricy*, *Extent*, *Solidity*, *Roundness*, dan *ShapeFactor4* memiliki *skewness* negatif. *Leptokurtic* hanya terjadi pada variabel: Perimeter dan EquivDiameter.



Gambar 4. Histogram tiap variabel

3.2. Model Training

Pada tahapan ini dilakukan fase training dengan menggunakan model training *Random Forest*, *Gradient Boosting Machine,* dan *Light* GBM (LGB). Pembagian data latih dan validasi yang baik dengan membagi data menjadi tiga yakni porsi dari data latih 70%, 80 %, dan 90%.

Tabel 3.Akurasi dan data latih

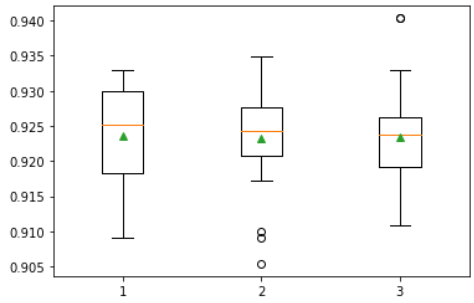
Persentase	Model	Akurasi
70 persen	RF	0.917184
70 persen	GBM	0.919074
70 persen	LGB	0.918758
80 persen	RF	0.917626
80 persen	GBM	0.918076
80 persen	LGB	0.918076
90 persen	RF	0.917626
90 persen	GBM	0.917789
90 persen	LGB	0.919340

Sembilan data yang diperoleh hasil pelatihan, didapatkan bahwa rata-rata akurasi berada pada nilai 0.91. Namun nilai akurasi terbaik menggunakan LGB. Pelatihan dilakukan menggunakan *default* model pelatihan tanpa ada penambahan *hyperparameter*.

3.3. *BoxCox Repeated* k-folds

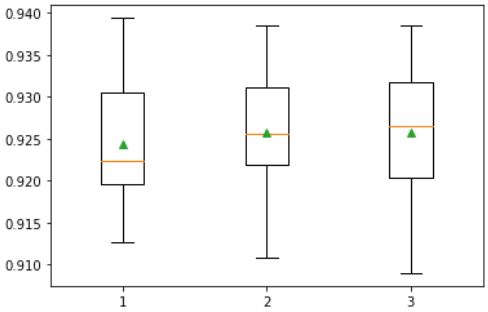
Pada tahapan ini, dataset divalidasi menggunakan *repeated* 10-folds sebanyak 3 kali. Dataset tidak dilakukan transformasi menggunakan *BoxCox*. Pada Gambar 5, dapat dilihat bahwa pada pengulangan pertama didapatkan bahwa rata-rata berada di 0.925 dan *quartile* 1 tidak mencapai 0.935. Terjadi penurunan pada pengulangan ke-tiga, walau tidak terlalu signifikan. Dari pengulangan tiga kali, tidak terdapat quartile 1 mencapai akurasi 93,5 persen.

DOI: https://doi.org/10.29207/resti.v5i1.3682



Gambar 5. Sebelum *BoxCox* 3 Repeated 10-folds.

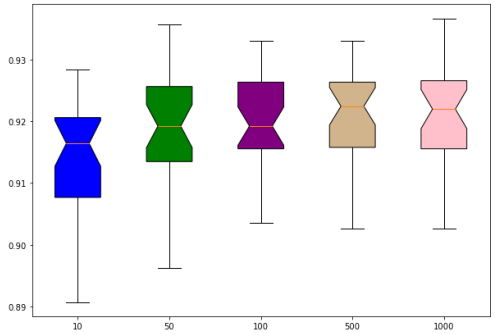
Penerapan transformasi *BoxCox* pada dataset ternyata dapat meningkatkan akurasi dari klasifikasi. Pada Gambar 6, terlihat bahwa hampir semua validasi 10-*folds* menunjukkan nilai quartile 1 yang mencapai 94 %. Dimana terjadi kenaikan nilai rata-rata akurasi setiap terjadi pengulangan validasi.



Gambar 6. Setelah *BoxCox* 3 Repeated 10-folds.

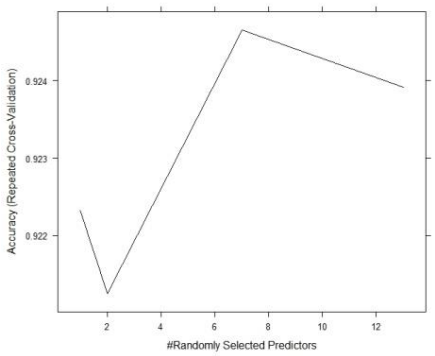
3.4. Random Forest

Salah satu parameter untuk optimasi dari metode *Random Forest* adalah jumlah pohon. Pada uji latihan dengan jumlah pohon 10, 50, 100, 500, dan 1000 didapatkan akurasi rata-rata sebesar 91,4, 92, 92, 92,1, dan 92,2 persen. Dimana nilai akurasi stabil pada jumlah pohon lebih dari 50 buah. Hubungan antara akurasi dan jumlah pohon dapat dilihat pada Gambar 7.



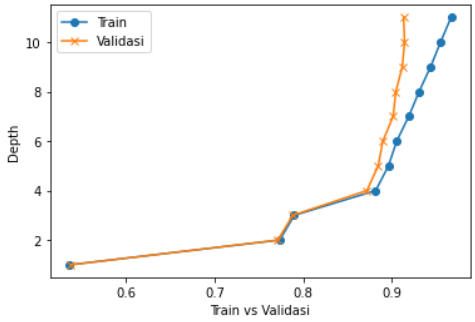
Gambar 7. Akurasi vs Jumlah Pohon

Jumlah prediktor menurut beberapa penelitian dapat mempengaruhi hasil dari akurasi *random forest.* Untuk itu dari 16 prediktor dengan 10.891 baris dan 7 kelas, dilakukan seleksi acak terhadap prediktor tersebut. Didapatkan bahwa prediktor terbaik sebanyak 7 variabel. Dengan nilai yang tidak terlalu beda, hanya berbeda 0.001 – 0.003. Akurasi terbaik bernilai 92.46 persen. Hubungan antara akurasi dan prediktor dapat dilihat pada Gambar 8.



Gambar 8. Akurasi vs jumlah prediktor

Hasil training pertama menunjukkan jumlah pohon keputusan sebanyak 50 untuk tahapan latihan berikutnya. Pada tahapan ini dilakukan variasi parameter *depth* dari 1 hingga 11. Dari Gambar 9, dapat dilihat bahwa hasil akurasi train dan validasi yang sama-sama naik drastis terhenti pada *depth*=7, dimana nilai nya sebesar 0.919177 dan 0.901579. Pada *depth* ke 8 hingga 11, nilai akurasi train naik terus hingga 0.966936 , namun nilai akurasi validasi stabil di angka 0.91. Akurasi validasi tertinggi pada *depth* 10 yakni 0.914433. Setelah *depth* 11, nilai validasi turun. Nilai train dan validasi terendah berada di *depth* = 1 yakni hanya bernilai 0.535911 dan 0.537642. Dengan demikian, maka untuk prediksi klasifikasi diambil parameter jumlah pohon 50 dan *depth* 10.



Gambar 9. RF dengan variasi *depth*.

Tabel 4, diperoleh akurasi tertinggi pada kacang kelas Bombay dengan presisi 1.00 dan *recall* 1.00. Hal ini dikarenakan jumlah kelas Bombay paling kecil, sehingga model lebih mudah dalam melakukan prediksi. Walaupun demikian, berdasarkan pada Tabel 7, volume kelas tertinggi terdapat pada Dermason tapi nilai akurasi nya mencapai 0,92 tidak begitu jauh dengan hasil kelas yang lain. Sedangkan yang terburuk terdapat pada kelas Sira. Sedangkan nilai sensitivitas (*recall*) terendah berada pada Barbuya yakni 0.88 persen , dimana nilai FN nya bernilai besar. Disisi lain, presisinya bernilai 0.94.

DOI: https://doi.org/10.29207/resti.v5i1.3682

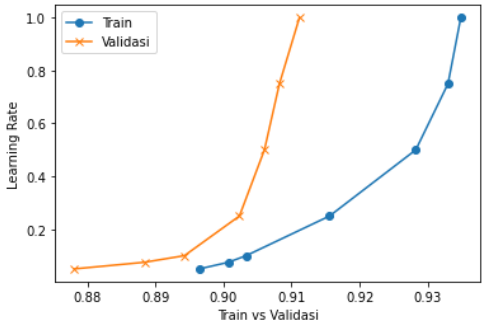
Tabel 4.Nilai Presisi, Recall dan f1-score RF

Kelas	Presisi	Recall	f1-score
Barbunya	0.94	0.88	0.91
Bombay	1.00	1.00	1.00
Cali	0.92	0.96	0.94
Dermason	0.92	0.90	0.91
Horoz	0.95	0.95	0.95
Seker	0.93	0.92	0.93
Sira	0.83	0.86	0.84

Didapatkan juga bahwa nilai kappa sebesar 0.89. Nilai ini mendekati 1, menandakan bahwa nilai koefisiennya menunjukan terdapat korelasi. Tabel 7 dan 8, pada confusion matrix *Random Forest* dan GBM memiliki hasil yang mirip, yakni pada hampir semua kesalahan yang terjadi. Kesalahn tersebar dalam prediksi terjadi pada kelas kacang Sira.

3.5. GBM

Pelatihan GBM dengan parameter *learning rate* (lr) digunakan untuk optimasi model training. Variasi nilai learning rate berkisar antara 0 -1. Gambar 10 tampak bahwa nilai validasi lebih rendah dari nilai *train* pada lr 0.05. Terdapat perbedaan yang tajam pada lr 1.0, dimana nilai *train* sebesar 0.93 dan nilai validitas 0.91.



Gambar 10. Learning Rate pada GBM

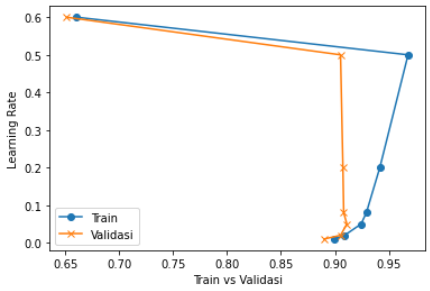
Pada Tabel 5, diperoleh kembali akurasi tertinggi pada kacang kelas Bombay dengan presisi 1.00 dan *recall* 1.00. Berdasarkan Tabel 8, volume kelas tertinggi terdapat pada Dermason tapi nilai akurasi nya sama dengan prediksi pada RF yakni 0,92, namun *recall* (sensivitas) nya lebih rendah. Sedangkan yang terkecil, masih terdapat pada kelas Sira. Sedangkan nilai sensitivitas (*recall*) pada Barbuya naik senilai 92 persen, dimana tersebut lebih besar dari model RF.

Tabel 5. Nilai Presisi, Recall dan f1-score GBM

Kelas	Presisi	Recall	f1-score
Barbunya	0.91	0.92	0.92
Bombay	1.00	1.00	1.00
Cali	0.94	0.95	0.95
Dermason	0.92	0.89	0.90
Horoz	0.94	0.95	0.95
Seker	0.93	0.93	0.93
Sira	0.83	0.86	0.84

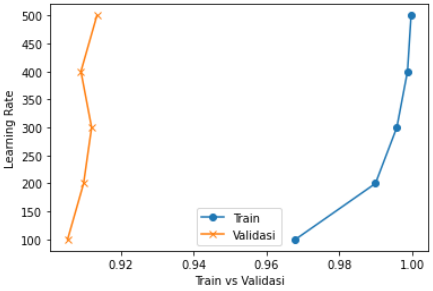
3.6. *Light* GBM

Fase training LGB, digunakan *learning rate* yang bervariasi antara 0 hingga 1. Nilai lr yakni 0.01, 0.02, 0.05, 0.08, 0.2, 0.5, dan 0.6. Terlihat pada Gambar 11, tampak bahwa terjadi penurunan tajam akurasi *train* dan validitas pada lr 0.6. Sebelumnya nilai train sebesar 0.967 pada lr 0.5 langsung turun menjadi 0.66 dan juga nilai validasi dari 0.90 menjadi 0.65. Nilai validitas tertinggi pada lr 0.20 yakni sebesar 0.907.



Gambar 11. *Learning Rate* pada LGB

Proses ini mendapatkan nilai lr terbaik yakni 0.2, maka fase training dilanjutkan dengan menggunakan parameter n_estimator. Penelitian ini didapatkan hasil yang signifikan berbeda antara nilai akurasi *train* dan validasi. Seperti tampak pada Gambar 12, bahwa pada n_estimator 100 nilai akurasi *train* dan validasi yakni 0.96 dan 0.90. Didapatkan bahwa fase *train* hampir mencapai angka satu yakni bernilai 0.999755 pada n_estimator 500. Namun pada akurasi validasi, nilai nya hanya berkisar 0.91.



Gambar 12. n_estimator pada LGB

Tampak pada Tabel 6, nilai akurasi pada tiap kelas tidak mengalami perbedaan signifikan pada model training sebelumnya yakni RF dan GBM. Pada Tabel 9, terlihat kesalahan dalam menentukan kelas kacang tertinggi ada pada kelas kacang Sira.

Tabel 6. Nilai Presisi, Recall dan f1-score GBM

Kelas	Presisi	Recall	f1-score
Barbunya	0.91	0.90	0.91
Bombay	1.00	1.00	1.00
Cali	0.92	0.95	0.93
Dermason	0.92	0.90	0.90
Horoz	0.95	0.94	0.94
Seker	0.94	0.92	0.93
Sira	0.84	0.86	0.85

DOI: https://doi.org/10.29207/resti.v5i1.3682

Tabel 7. Confusion Matrix Random Forest

Confusion	Barbuya	Bombay	Cali	Dermason	Horoz	Seker	Sira
Barbunya	133	0	8	0	2	0	1
Bombay	0	33	0	0	0	0	0
Cali	5	0	164	0	1	0	2
Dermason	1	0	0	333	2	7	33
Horoz	3	0	1	2	184	0	4
Seker	3	0	0	4	0	171	6
Sira	1	0	1	22	6	6	223

Tabel 8. Confusion Matrix Gradient Boosting Machine

Confusion	Barbuya	Bombay	Cali	Dermason	Horoz	Seker	Sira
Barbunya	130	0	12	0	0	0	2
Bombay	0	33	0	0	0	0	0
Cali	5	0	163	0	2	0	2
Dermason	0	0	0	340	2	4	30
Horoz	4	0	2	1	182	0	5
Seker	3	0	0	7	0	170	4
Sira	1	0	1	21	6	6	224

Tabel 9. Confusion Matrix Light GBM

Confusion	Barbuya	Bombay	Cali	Dermason	Horoz	Seker	Sira
Barbunya	127	0	12	0	2	0	3
Bombay	0	33	0	0	0	0	0
Cali	2	0	165	0	3	0	2
Dermason	0	0	0	338	0	6	32
Horoz	4	0	0	2	184	0	4
Seker	1	0	1	6	0	170	6
Sira	1	0	2	23	5	6	222

4. Kesimpulan

Penggunaan model transform *BoxCox* dapat meningkatkan akurasi dari klasifikasi, walau perbedaan secara nilai tidak terlalu besar jika tanpa menggunakan *BoxCox*. Untuk prediksi klasifikasi dengan nilai *False* Negatif terkecil terdapat pada model training GBM, walau demikian nilai *recall* secara rata-rata tidak terlalu signifikan perbedaan dengan model lain. Jika ingin mendapatkan prediksi dengan salah, artinya lebih sedikit salah dalam memprediksi kelas dari kacang.

RF, GBM, dan LGB memiliki hasil prediksi klasifikasi kelas yang hampir sama. Namun dalam fase training, memiliki perbedaan yang cukup signifikan. RF memiliki kemampuan akurasi training hingga 96%, GBM hanya mampu pada angka 93 persen, sedangkan LGB yang awalnya terjadi penurunan tajam pada parameter *learning rate* namun pada parameter estimator mampu mencapai 99,9 persen. Nilai akurasi pada tiap kelas berbeda, namun prediksi terbaik terjadi pada kelas kacang Bombay sedangkan prediksi terburuk terjadi pada kelas kacang Sira. Proses prediksi ini didapatkan hasil bahwa kelas kacang Barbunya, Bombay, Cali, Dermason, Horoz, Seker, dan Sira memberikan nilai akurasi berturut-turut yakni 91%, 100%, 92%, 92%, 95%, 94% dan 84%.

Daftar Rujukan

[1] D. F. Barbin et al., "Classification and compositional characterization of different varieties of cocoa beans by near infrared spectroscopy and multivariate statistical analyses," J. Food Sci. Technol. 2018 557, vol. 55, no. 7, pp. 2457-2466, Apr. 2018. https://doi.org/10.1007/s13197-018-3163-5

[2] R. Megías-Pérez, S. Grimbs, R. N. D'Souza, H. Bernaert, and N. Kuhnert, "Profiling, quantification and classification of cocoa beans based on chemometric analysis of carbohydrates using hydrophilic interaction liquid chromatography coupled to mass spectrometry," Food Chem., vol. 258, pp. 284-294, Aug. 2018. https://doi.org/10.1016/j.foodchem.2018.03.026

[3] A. J. Myles, S. D. Brown, and T. A. Zimmerman, "Transfer of Multivariate Classification Models Between Laboratory and Process Near-Infrared Spectrometers for the Discrimination of Green Arabica and Robusta Coffee Beans," Appl. Spectrosc. Vol. 60, Issue 10, pp. 1198-1203, vol. 60, no. 10, pp. 1198-1203, Oct. 2006. https://doi.org/10.1366/000370206778664581

[4] F. Kurniawan, I. W. Budiastra, Sutrisno, and S. Widyotomo, "Classification of Arabica Java Coffee Beans Based on Their Origin using NIR Spectroscopy," IOP Conf. Ser. Earth Environ. Sci., vol. 309, no. 1, p. 012006, Sep. 2019. https://doi.org/10.1088/1755-1315/309/1/012006

[5] A. Vázquez-Ovando, F. Molina-Freaner, J. Nuñez-Farfán, D. Betancur-Ancona, and M. Salvador-Figueroa, "Classification of cacao beans (Theobroma cacao L.) of southern Mexico based on chemometric analysis with multivariate approach," Eur. Food Res. Technol. 2015 2406, vol. 240, no. 6, pp. 1117-1128, Feb. 2015. https://doi.org/10.1007/s00217-015-2415-0

[6] Y. Li, J. Sun, X. Wu, Q. Chen, B. Lu, and C. Dai, "Detection of viability of soybean seed based on fluorescence hyperspectra and CARS-SVM-AdaBoost model," J. Food Process. Preserv., vol. 43, no. 12, p. e14238, Dec. 2019. https://doi.org/10.1111/jfpp.14238

[7] M. Koklu and I. A. Ozkan, "Multiclass classification of dry beans using computer vision and machine learning techniques," Comput. Electron. Agric., vol. 174, p. 105507, Jul. 2020. https://doi.org/10.1016/j.compag.2020.105507

DOI: https://doi.org/10.29207/resti.v5i1.3682

[8] S. A. Araújo, W. A. L. Alves, P. A. Belan, and K. P. Anselmo, "A Computer Vision System for Automatic Classification of Most Consumed Brazilian Beans," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9475, pp. 45-53, Dec. 2015. https://doi.org/10.1007/978-3-319-27863-6_5

[9] E. M. De Oliveira, D. S. Leme, B. H. G. Barbosa, M. P. Rodarte, and R. G. F. Alvarenga Pereira, "A computer vision system for coffee beans classification based on computational intelligence techniques," J. Food Eng., vol. 171, pp. 22-27, Feb. 2016. https://doi.org/10.1016/j.jfoodeng.2015.10.009

[10] F. A. Santos, A. M. P. Canuto, B. R. C. Bedregal, E. S. Palmeira, and I. N. P. Silva, "Supervised Methods Applied to the Construction of a Vision System for the Classification of Cocoa Beans in the Cut-Test," An. do Encontro Nac. Inteligência Artif. e Comput., pp. 72-83, Oct. 2019.

[11] A. J. Ona Ona, F. Grijalva, K. Proano, B. Acuna, and M. Garcia, "Classification of fresh cocoa beans with pulp based on computer vision," 2020 IEEE ANDESCON, ANDESCON 2020, Oct. 2020. https://doi.org/10.1109/ANDESCON50619.2020.9272188

[12] R. M. Sakia, "The Box-Cox transformation technique: a review," J. R. Stat. Soc. Ser. D …, 1992. https://doi.org/10.2307/2348250

[13] M. J. Gurka, L. J. Edwards, K. E. Muller, and L. L. Kupper, "Extending the Box-Cox transformation to the linear mixed model," J. R. Stat. Soc. Ser. A (Statistics Soc., vol. 169, no. 2, pp. 273-288, Mar. 2006. https://doi.org/10.1111/j.1467-985X.2005.00391.x

[14] J. Osborne, "Improving your data transformations: Applying the Box-Cox transformation," Pract. Assessment, Res. …, 2010.

[15] M. Z. Hossain, "The use of Box-Cox transformation technique in economic and statistical analyses," J. Emerg. Trends Econ. …, 2011.

[16] L. Wang, X. Zhou, X. Zhu, Z. Dong, and W. Guo, "Estimation of biomass in wheat using random forest regression algorithm and remote sensing data," Crop J., vol. 4, no. 3, pp. 212-219, Jun. 2016. https://doi.org/10.1016/j.cj.2016.01.008

[17] A. T. Azar, H. I. Elshazly, A. E. Hassanien, and A. M. Elkorany, "A random forest classifier for lymph diseases," Comput. Methods Programs Biomed., vol. 113, no. 2, pp. 465-473, Feb. 2014. https://doi.org/10.1016/j.cmpb.2013.11.004

[18] S. Vitrack-Tamam et al., "Random Forest Algorithm Improves Detection of Physiological Activity Embedded within Reflectance Spectra Using Stomatal Conductance as a Test Case," Remote Sens. 2020, Vol. 12, Page 2213, vol. 12, no. 14, p. 2213, Jul. 2020. https://doi.org/10.3390/rs12142213

[19] S. Agajanian, O. Oluyemi, and G. M. Verkhivker, "Integration of random forest classifiers and deep convolutional neural networks for classification and biomolecular modeling of cancer driver mutations," Front. Mol. Biosci., vol. 6, no. JUN, p. 44, 2019. https://doi.org/10.3389/fmolb.2019.00044

[20] C. L. Eng, J. C. Tong, and T. W. Tan, "Predicting host tropism of influenza A virus proteins using random forest," BMC Med. Genomics, vol. 7, no. 3, pp. 1-11, Dec. 2014. https://doi.org/10.1186/1755-8794-7-S3-S1

[21] S. Touzani, J. Granderson, and S. Fernandes, "Gradient boosting machine for modeling the energy consumption of commercial buildings," Energy Build., vol. 158, pp. 1533-1543, Jan. 2018. https://doi.org/10.1016/j.enbuild.2017.11.039

[22] Y. Zhang and A. Haghani, "A gradient boosting method to improve travel time prediction," Transp. Res. Part C Emerg. Technol., vol. 58, pp. 308-324, Sep. 2015. https://doi.org/10.1016/j.trc.2015.02.019

[23] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," Artif. Intell. Rev. 2020 543, vol. 54, no. 3, pp. 1937-1967, Aug. 2020. https://doi.org/10.1007/s10462-020-09896-5

[24] G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," Adv. Neural Inf. Process. Syst., vol. 30, 2017.

[25] X. Ma, J. Sha, D. Wang, Y. Yu, Q. Yang, and X. Niu, "Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning," Electron. Commer. Res. Appl., vol. 31, pp. 24-39, Sep. 2018. https://doi.org/10.1016/j.elerap.2018.08.002

[26] D. Wang, Y. Zhang, and Y. Zhao, "LightGBM: An effective miRNA classification method in breast cancer patients," ACM Int. Conf. Proceeding Ser., pp. 7-11, Oct. 2017. https://doi.org/10.1145/3155077.3155079

[27] "UCI Machine Learning Repository: Dry Bean Dataset Data Set." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Dry+Bean+Dataset. [Accessed: 26-Nov-2021].

[28] M. Pal, "Random forest classifier for remote sensing classification," http://dx.doi.org/10.1080/01431160412331269698, vol. 26, no. 1, pp. 217-222, Jan. 2007. https://doi.org/10.1080/01431160412331269698

DOI: https://doi.org/10.29207/resti.v5i1.3682