Cenderawasih : Journal of Statistics and Data Science
E-ISSN: 3026-0981

PERBANDINGAN KINERJA MODEL LINEAR DISCRIMINANT ANALYSIS DAN DECISSION TREE PADA STUDI KASUS DATASET AUDIT DATA

FEBRYANA D. HANAFI[1], WAHYU APRILLIA S.[2], HALLE F. P. WATORY[3], MUHAMMAD ASGHAR NAZAL[4]
[1,2,3,4)]Program Studi Statistika Fakultas MIPA Universitas Cenderawasih Jayapura, Indonesia
e-mail: waprilliasa@gmail.com

ABSTRAK

Penelitian ini akan membandingkan dua metode penambangan data yang umum digunakan, yaitu *Linear Discriminant Analysis* (LDA) dan *Decission Tree* dengan nilai *accuracy* 99,6% memiliki kinerja lebih baik dan relatif optimal pada *dataset audit data*. Dengan tujuan mengoptimalkan analisis statistik *audit data*, pengujian ini mengidentifikasi model mana yang lebih baik dan relatif optimal pada *dataset audit data*. Penelitian dilakukan untuk membandingkan performa *machine learning* dengan memilih subkumpulan fitur yan relevan dari data. Perbandingan dengan menggunnakan nilai *accuracy* menunjukkan jika model *Decission Tree* dan hasil *accuracy* sebesar 99,6%.

Kata Kunci: Linear Discriminant Analysis, Decission Tree, Audit Data, Accuracy

1. PENDAHULUAN

Di era digital yang berkembang pesat, produksi dan pengumpulan data telah mencapai proporsi yang belum pernah terjadi sebelumnya. Organisasi dan bisnis dihadapkan dengan semakin banyaknya data dari berbagai sumber, termasuk transaksi bisnis, sensor, media sosial, perangkat seluler, dan banyak lagi. Dalam konteks ini, *data mining* sebagai metode analisis berperan penting dalam mengubah data menjadi wawasan yang berharga.

Teknik data mining dapat digunakan untuk menemukan informasi berguna dalam kumpulan data besar.Data mining merupakan bidang keilmuan yang relatif baru dengan beragam penerapan, menjadikannya salah satu dari sepuluh bidang keilmuan dengan dampak terbesar terhadap perkembangan teknologi (Ghorbani & Ghousi, 2019). *Data mining* adalah proses mengekstraksi pola, informasi, dan pengetahuan yang berguna dari data yang besar dan kompleks (Han et al, 2011).

Ini menggunakan berbagai teknik dan algoritma yang dirancang untuk mengidentifikasi hubungan, tren, dan anomali dalam data. *Data mining* memungkinkan organisasi untuk mengambil keputusan yang lebih informasional, mengoptimalkan operasional, meningkatkan kepuasan pelanggan, dan bahkan mengantisipasi perubahan pasar. Ada berbagai macam model yang cukup populer, diantaranya yaitu *Decision Trees*, *Naïve Bayes Classifiers*, *Neural Networks*, *Statistical Analysis*, *Genetic Algorithms*, *Rough Sets*, *K-Nearest Neighbor Classifier*, *Rule-based Methods*, *Memory Based Reasoning*, dan juga *Support Vector Machines* (Gorunescu, 2011).

Terdapat berbagai jenis teknik *data mining* telah dikembangkan dan diterapkan pada sejumlah *dataset*. Meskipun begitu cakupan dan fokusnya tidaklah selalu konsisten, sehingga kinerja dari teknik ini selalu bervariasi dari satu permasalahan ke permasalahan lainnya. Maka dari itu, peneliti perlu mengeksplorasi model dari *data mining* yang relatif optimal untuk segera menyelesaikan masalah yang mereka hadapi dalam sebuah penelitian pada suatu *dataset* (Zhang et al, 2017).

Banyak penelitian telah dilakukan untuk membandingkan algoritma klasifikasi, diantaranya telah dilakukan oleh Fatmawati (2016) dengan hasil penelitiannya yaitu model *Naïve Bayes* dengan *accuracy* 75,13%

lebih baik dibandingkan model *C4.5* dengan *accuracy* 73,30%. Sedangkan penelitian yang dilakukan oleh Marcos dan Utomo (2015) dengan membandingkan model yang sama dengan yang dilakukan Fatmawati (2016), yaitu model *Naïve Bayes* dan model *C4.5* dengan perolehan nilai *accuracy* secara berurutan, yaitu 77,47% dan 74,87%. Dari hasil kedua peneltian tersebut digunakan *dataset* yang sama, namun nilai *accuracy* dari penelitian yang dilakukan oleh Marcos dan Utomo (2015) lebih baik dikarenakan adanya perlakuan seleksi fitur yang menghasilkan atribut yang digunakan lebih sedikit.

Yusuf dan Khadijah (2022) melakukan perbandingan model *Support Vector Machine*, *K-Nearest Neighbor*, *Naïve Bayes*, dan *Logistic Regression* dengan menghasilkan nilai *accuracy* secara urutan sebesar 69,15%, 67,45%, 66,60%, dan 53,62%. Dari semua model yang diterapkan, model *SVM* memiliki nilai *accuracy* tertinggi mengartikan jika model *SVM* dapat melakukan klasifikasi yang lebih optimal jika dibandingkan model lainnya.

Permasalahan yang cukup kompleks ini akan dihadapi dalam pemilihan model optimal yang akan diterapkan pada klasifikasi *data mining*. Terdapat 2 cara yang dilakukan dalam penentuan model yang akan digunakan, yaitu dengan memilih model berdasarkan kelas tertentu dan juga memilih model secara bebas sesuai dengan kehendak peneliti (Gorunescu, 2011). Namun pada umumnya seperti studi kasus diatas, penelitian dilakukan dengan memilih secara bebas model yang digunakan dalam klasifikasi *data mining*, hal yang sama dilakukan pada penelitian ini.

Dalam kaitannya dengan studi kasus ini, fokus utama adalah penerapan *data mining* dalam analisis *audit data*. *Audit data* merupakan data yang membutuhkan perhatian khusus dalam pengelolaan data bisnis. *Audit data* adalah proses penting yang memastikan *accuracy*, kepatuhan, dan efisiensi dalam operasi organisasi. Dengan tujuan mengoptimalkan analisis *audit data*, penelitian ini akan membandingkan dua metode penambangan data yang umum digunakan, yaitu *Linear Discriminant Analysis* (LDA) dan *Decission Tree* (DT). Metode LDA merupakan algoritma statistik yang bertujuan untuk memaksimalkan pemisahan antar kelompok data, sedangkan *Decission Tree* merupakan model yang menggambarkan hubungan antar variabel dalam data.

Tujuan utama studi kasus ini adalah membandingkan kinerja metode *Linear Discriminant Analysis* dan *Decision Tree* dalam konteks audit dan data risiko. Melalui perbandingan ini, peneliti berharap dapat memberikan wawasan yang lebih luas mengenai efektivitas masing-masing metode dalam menangani data audit serta tantangan dan tujuan risiko. Dengan kemajuan teknologi yang berkelanjutan dan akses terhadap data yang terus meningkat, penting untuk lebih memahami bagaimana *data mining* dapat diterapkan dalam konteks bisnis. Hasil penelitian ini diharapkan dapat memberikan panduan praktis bagi organisasi untuk mengambil keputusan berdasarkan data dan meningkatkan manajemen risiko mereka.

2. METODE PENELITIAN

Beberapa tahapan yang dilakukan dalam penelitian ini, diantaranya yaitu pengumpulan data, *pre-processing* data, dan perbandingan model.

2.1 Sumber Data

Dataset yang digunakan merupakan data tentang perusahaan yang diambil dari *repository UCI Learning Machine*. *Audit data* merupakan proses pemeriksaan, evaluasi, dan verifikasi sistem, proses, dan data yang ada dalam suatu organisasi atau sistem. Dimana tujuan dari *audit data* sendiri untuk memastikan keakuratan, keandalan, keamanan, dan ketersediaan data yang digunakan. Dalam *audit data* diperkirakan mengenai data transaksi, data keuangan, data pelanggan, dan data lainnya yang relevan. Selain itu, *audit data* dapat membantu meningkatkan efisiensi operasional, mengurangi risiko, dan memastikan kepatuhan terhadap peraturan dan standar yang berlaku.

2.2 Pre-processing Data

Pre-processing merupakan proses mempersiapkan data mentah sebelum dilakukan proses lainnya pada data. Pada dasarnya *pre-processing* dilakukan dengan melakukan eliminasi atau pengahpusan data yang tidak sesuai atau mengubah data menjadi format yang lebih mudah untuk diproses oleh sistem. Pada *pre-processing* ini sangat penting dilakukan dalam analisis sentimen, khususnya pada media sosial. Sebaian besar diantaranya berisikan kata- kata ataupun kalimat informal dan tidak terstruktur sehingga memiliki *noise* yang cenderung besar (Clark, 2003).

2.2.1 Data Cleaning

Pada proses ini dilakukan modifikasi atau penghapusan data yang dianggap tidak akiurat atau tidak sesuai, duplikat data, data tidak lengkap, kesalahan format, dan juga rusak dalam kumpulan data yang dimiliki yang nantinya akan diproses. Adapun beberapa tahapan yang dilakukan pada proses *data cleansing*, yaitu:

1. *Duplicate data*
   Pada *dataset audit data* terdeteksi sebanyak 13 data yang duplikat, sehingga perlu dilakukannya penghapusan data yang duplikat. Dari proses pengahpusan data duplikat yang terdeteksi, terdapat perubahan pada baris data dari 776 menjadi 763 data.
2. *Missing value*
   Proses mendeteksi nilai yang hilang pada suatu *dataset*. Pada *dataset audit data* terdeteksi adanya 1 missing value pada atribut *Money_Value*.
3. *Imputation missing value*
   Merupakan proses yang digunakan untuk mengisi nilai yang hilang pada atribut yang terdeteksi pada tahap *missing value*. Untuk mengatasi nilai yang hilang pada *dataset* yang memiliki atribut numerik, dapat digunakan metode penggantian nilai hilang dengan mencari *median*. Menggunakan *median* lebih efektif daripada menggunakan *mean*, karena akan mengurangi pengaruh pada deteksi *outlier*. Namun, pada umumnya keduanya memberikan hasil imputasi yang seimbang.
4. *Outlier*
   Outlier adalah data di luar kisaran nilai normal yang dapat mempengaruhi hasil dan kesimpulan statistik. Penting untuk mengidentifikasi dan mempertimbangkan sifat dan penyebab *outlier* sebelum menanganinya.

2.2.2 Transformasi Data

Transformasi data adalah proses pengubahan data dari skala pengukuran data asli ke bentuk skala lain dengan tujuan untuk memperbaiki karakteristik data dan memudahkan analisis. Transformasi data dapat dilakukan dengan berbagai cara, diantaranya dengan mengubah satuan ukuran data, mengubah sebaran data, atau mengubah bentuk data. Pada studi kasus ini dilakukan transformasi data dengan menggunakan metode *Z-Score normalization*. Dimana *Z-Score normalization* merupakan sebuah metode normalisasi yang digunakan dalam statistik dan analisis data sebagai pengubah variabel numerik dalam sebuah *dataset* menjadi distribusi normal dengan rata-rata atau *mean* 0 (nol) dan standar deviasinya 1 (satu). Metode ini digunakan dalam memperbaiki skala data dan memudahkan perbandingan antara berbagai variabel yang memiliki skala berbeda. Pada *dataset audit data* didapati Y (hasil) data merupakan data pada atribut *Risk* yang merupakan hasil klasifikasi dari hasil-hasil atribut lainnya.

2.2.3 Featur Selection

Feature selection bertujuan meningkatkan performa model pembelajaran mesin dengan memilih subkumpulan fitur yang relevan dari data. Hal ini dilakukan dengan mengurangi jumlah variabel yang dipilih dari variabel yang ada tanpa membentuk variabel baru.

2.2.4 Wrapper Method

Metode *wrapper* adalah metode pemilihan fitur dalam *machine learning* yang menggunakan model pembelajaran untuk mengeluarkan setiap subset fitur dan memilih subset fitur yang optimal. Teknik ini secara berulang memilih subkumpulan fitur dan membangun model pembelajaran pada setiap iterasi. Pada setiap iterasi, subset fitur yang dihasilkan dievaluasi berdasarkan performa model pembelajaran yang dibuat. Proses ini berlanjut hingga ditemukan subkumpulan fitur terbaik yang memberikan performa model terbaik. Ada berbagai jenis metode *wrapper* dianataranya yaitu metode *stepwise* yang digunakan pada *dataset* ini. Metode *stepwise* digunakan untuk menentukan model regresi terbaik dengan memilih variabel signifikan dalam memprediksi variabel target dalam analisis statistik.

2.2.5 Split Data

Proses ini merupakan proses yang membagi *dataset* menjadi 2 *subset* yang berbeda, yaitu *training set* dan *testing set*. *Training set* digunakan sebagai bentuk pelatihan model pada *machine learing*, sedangkan *testing set* digunakan sebagai bentuk pengujian kinerja model yang telah dibor. Adapun tujuan dari *split data*, yaitu untuk menghindari adanya *overfitting* dan memastikan jika model yang dibangun dapat digeneralisasikan dengan baik pada sebuah data. Pada *split data* digunakan metode *holdout*, yaitu dengan membagi *dataset* dengan proporsi tertentu. Pada *dataset* ini proporsi yang digunakan adalah 20:80 untuk *testing set* dan *training set*.

2.2.6 Imbalance Data

Imbalance data merupakan masalah umum yang sering terjadi pada *machine learning*. Hal ini dikarenakan jumlah sampel yang terkait dengan setiap kelas sangat beragam. Jika masalah ini terus terjadi akan mengakibatkan performa model yang buruk dan juga masalah pada korelasi atribut, pemisahan kelas, dan juga penilaian. Dalam penanganan masalah ini penting untuk memastikan jika model dapat menggeneralisasikan data yang tidak terdeteksi dengan baik dan menghindari *overftting*. Pada *dataset* ini digunakan metode *Synthetic Minority Over-sampling Technique* (SMOTE). Metode SMOTE seringkali digunakan untuk menghasilkan sampel sintetis. Ini akan melibatkan pembuatan sampel baru dengan melakukan interpolasi antara sampel yang ada pada kelas minoritas.

2.3 Model Kinerja

2.3.1 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) adalah bentuk umum dari diskriminan linier *Fisher*, metode yang digunakan dalam ilmu statistik, pengenalan pola, dan pembelajaran mesin untuk menemukan asosiasi. Linearitas fitur menggambarkan atau memisahkan dua atau lebih objek atau peristiwa. Kombinasi yang dihasilkan dapat digunakan sebagai pengklasifikasi linier atau biasa digunakan untuk reduksi dimensi sebelum klasifikasi. Tujuan dari analisis *Linear Discriminant Analysis* (LDA) adalah untuk mengklasifikasikan objek ke dalam beberapa kelas berdasarkan karakteristik yang 15 mendeskripsikan objek tersebut. Pada analisis diskriminan linier, objek mempunyai dua variabel yaitu variabel kelas/variabel ikutan (variabel terikat) dan atribut/bebas (variabel bebas), variabel terikat mempunyai yang berhubungan dengan variabel bebas menjelaskan variabel ini (Wei dan Tan, 2022).

Analisis diskriminan adalah teknik analisis ketergantungan statistik yang berguna untuk mengklasifikasikan beberapa kelompok objek. Kelompok analisis diskriminan ini terjadi karena pengaruh satu atau lebih variabel lain yang merupakan variabel bebas. Fungsi diskriminan terbentuk dari kombinasi linier variabel (Tatham et. Al., 1998). Adapun beberapa tujuan dari analisis *Linear Discriminant Analysis* (LDA), antara lain sebagai berikut:

1. Menentukan apakah terdapat perbedaan yang jelas antar kelompok.
2. Klasifikasikan benda dan cari tahu apakah benda termasuk golongan 1 atau golongan 2 atau golongan lain.
3. Tentukan seberapa akurat sistem dalam mengidentifikasi objek berdasarkan karakteristik tertentu.

Implementasi *Linear Discriminant Analysis* (LDA) ke dalam sebuah persamaan adalah sebagai berikut:

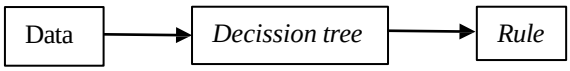
$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n$$

Keterangan:

Y = nilai (skor) diskriminan dan merupakan variabel terikat.
X_n = variabel (atribut) ke-n dan merupakan variabel bebas.
B_n = koefisien diskriminan/bobot dari variabel (atribut) ke-n.

2.3.2 Decission Tree

Decission tree adalah klasifikasi dan algoritma regresi yang merupakan bagian dari grup *ensemble learning*. *Decission tree* merupakan metode yang dapat digunakan untuk mengklasifikasikan tim objek atau data untuk menghasilkan sebuah keputusan (Achmad, 2012). Proses pada *Decission tree* adalah mengubah bentuk data tabel menjadi sebuah model pohon. Model pohon ini akan menghasilkan *rule* dan disederhanakan (Basuki & Syarif, 2003).



Gambar 1. Konsep *Decission Tree*

Model *Desiccion Tree* prediksi menggunakan struktur berhirarki. Konsep dari *Decission Tree* adalah mengubah data menjadi *Decission Tree* dan aturan-aturan keputusan. Dengan manfaat utamanya yaitu mem-*break down process* pengambilan keputusan akan lebih mudah menemukan solusi permasalahan yang ada. *Decission Tree* terbentuk dari proses pemilahan rekursif biner pada suatu gugus data sehingga nilai variabel respon pada setiap gugus akan lebih homogen. Terdapat 3 jenis *node*, antara lain:

1. Akar, merupakan *node* teratas, yaitu tidak memiliki *input* sehingga mempunyai *output* lebih dari satu.
2. *Internal node*, hanya terdapat satu *input* sehingga mempunyai *output* minimal dua.
3. Daun, merupakan *node* akhir yang hanya terdapat satu *input* dan tidak memiliki *output* (simpul terminal).

2.4 Perbandingan Kinerja Model

2.4.1 Accuracy

Accuracy mengacu pada metrik evaluasi digunakan untuk mengukur sejauh mana model yang digunakan dapat memprediksi secara benar, kelas atau nilai target dari data yang tidak dikenal. *Accuracy* dihitung dengan membandingkan prediksi model dengan nilai sebenarnya dalam *dataset* pengujian. Perhitungan ini dapat di-*input* pada tabel *confusion matrix* sebagai berikut:

Tabel 1. *Confusion Matrix Actual*

Predictions	***Positive***	***Negative***
Positive	TP	FP
Negative	FN	TN

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Keterangan:

TP *(True Positive)* = Jumlah data yang benar diprediksi sebagai positif.
TN *(True Negative)* = Jumlah data yang benar diprediksi sebagai negatif.
FP *(False Positive)* = Jumlah data yang salah diprediksi sebagai positif.
FN *(False Negative)* = Jumlah data yang salah diprediksi sebagau negatif.

3. HASIL DAN PEMBAHASAN

Pengujian ini dilakukan untuk membandingkan kinerja model pada *Linear Discriminant Analysis* (LDA) dan *Decission Tree* dengan mengidentifikasi model mana yang lebih baik dan cenderung optimal pada *dataset audit data*. Pengujian ini dilakukan dengan menggunakan *python* versi 3.10.9 yang dijalankan pada sistem

operasi *Ms. Windows* 11 *Enterprise* 64-bit dengan spesifikasi *processor Intel® Celeron®* N4020 CPU @ 1.10GHz dan 4GB RAM. Pada pengujian ini diperoleh evaluasi model yang dapat diperoleh dengan melihat nilai *accuracy* yang dihasilkan pada hasil *code* yang dilakukan.

Tabel 2. Hasil *Confusion Matrix*

Model	TP	FP	FN	TN	*Accuracy*
Linear Discriminant Analysis	80	4	20	49	87,6%
Decission Tree	84	0	3	66	99,6%

Pada Tabel di atas terlihat hasil dari *code confusion matrix* pada model yang dibuat untuk menentukan kinerja berdasarkan nilai TP, FP, FN, dan TN yang diperoleh. Nilai *accuracy* pada Tabel 2 menunjukkan persentase *accuracy* dari *Linear Discriminant Analysis* (LDA) dan *Decisssion Tree*. Dari Tabel 2 terlihat bahwa nilai *accuracy* dari kedua model berada di atas 80%. Nilai *accuracy* dari model *Decission Tree* lebih besar dari *Linear Discriminant Analysis* yaitu dengan persentase akurasi sebesar 99,6%. Sedangkan nilai *accuracy* dari model *Linear Discriminant Analysis*, yaitu sebesar 89,6% .

4. KESIMPULAN DAN SARAN

Dari hasil pengujian dengan membagi *dataset* dengan *ratio test*:*train*, yaitu 20:80 dengan menggunakan dua model sebagai perbandingan, yaitu *Linear Discriminant Analysis* (LDA) dan *Decission Tree*. Hasil perbandingan dengan menggunakan nilai *accuracy* menunjukkan jika model *Decission Tree* dengan nilai *accuracy* 99,6% memiliki kinerja lebih baik dan relatif optimal dibandingkan model *Linear Discriminant Analysis* (LDA) dengan nilai *accuracy* 87,6% yang dijalankan pada *dataset audit data*. Pengujian ini masih sebatas membandingkan nilai *accuracy* dari dua model algoritma sederhana. Diharapkan pengujian lebih lanjut mencakup pengembangan dengan menggunakan beberapa model yang lebih beragam dan kurang umum digunakan untuk menemukan model yang lebih baik dan lebih optimal untuk digunakan pada sebuah *dataset*.

DAFTAR PUSTAKA

Achmad, B. D. M., Slamat, F,. & ITATS, F. T. I. (2012). Klasifikasi Data Karyawan Untuk Menentukan Jadwal Kerja Menggunakan Metode *Desiccion* Tree. Jurnal IPTEK, 16(1).

Ansori, Y., & Khadijah, F. H. H. (2022). Perbandingan Metode *Machine Learning* Dalam Analisis Sentimen *Twitter*. Jurnal Sistem Dan Teknologi Informasi. Jawa Timur: Universitas Islam Negeri Maulana Malik Ibrahim.

Basuki, A., & Iwan, S. (2003). *Decision Tree*. Surabaya: Politeknik Electronika Negeri.

Clark, A. (2003). *Pre-processing Very Noisy Text*. *Proceedings of Workshop on Shallow Processing of Large Corpora* (pp. 12- 22). *Lancaster*: *Lancaster University*.

Fatmawati. (2016). Perbandingan Algoritma Klasifikasi *Data Mining* Model *C4.5* Dan *Naive Bayes* Untuk Prediksi Penyakit Diabetes. Jurnal *Techno* Nusa Mandiri, XIII (1), 50–59.

Ghorbani, R., & Ghousi, R. (2019). *Predictive Data Mining Approaches In Medical Diagnosis*: *A Review Of Some Diseases Prediction*. *International Journal of Data and Network Science*, 3, 47–70. https://doi.org/10.5267/j.ijdns.2019.1.003.

Gorunescu, F. (2011). *Data Mining*: *Concepts, Models and Techniques* (Vol. 12). *Berlin*, *Heidelberg*: *Springer Berlin Heidelberg*. https://doi.org/10.1007/978-3-642-19721-5.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining*: *Concepts and Techniques* (*3rd*)., *Elsevier*.https://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management- Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan- Kaufmann-2011.pdf.

Hooda, N. (2018). *Audit Data*. https://archive.ics.uci.edu/dataset/475/audit+data.

Tatham, R.L., Hair, J.F, Anderson, R.E., dan Black, W.C., (1998), "*Multivariate Data Analysis*", *Prentice Hall, New Jersey*.

Wei, Y., Gu, K., & Tan, L. (2022). *A Positioning Method For Maize Seed Lasercutting Slice Using Linear Discriminant Analysis Based On Isometric Distance Measurement*. *Information Processing In Agriculture*, 9(2), 224-232.

Zhang, Y., Xin, Y., Li, Q., Ma, J., Li, S., Lv, X., & Lv, W. (2017). *Empirical Study Of Seven Data Mining Algorithms On Different Characteristics Of Datasets For Biomedical Classification Applications*. *Biomedical Engineering OnLine*, 16(1), 125. https://doi.org/10.1186/s12938-017-0416-x.