JIIC: JURNAL INTELEK INSAN CENDIKIA
https://jicnusantara.com/index.php/jiic
Vol : 2 No: 12, Desember 2025
E-ISSN : 3047-7824

Penerapan Model Klasifikasi Biner Menggunakan Regresi Logistik pada Dataset Kismis

Application of Binary Classification Model Using Logistic Regression on Raisin Dataset

Armand Cahya Nugraha[1*], Hasbi Firmansyah[2], Wahyu Asriyani[3], Riski Prasetyo Tulodo[4]
Universitas Pancasakti Tegal
Email: armancahyan@gmail.com[1], hasbifirmansyah@upstegal.ac.id[2], asriyani1409@gmail.com[3]*
Rizki.prasetyo.tulodo@gmail.com[4]

Article Info

Article history :
Received : 19-12-2025
Revised : 20-12-2025
Accepted : 22-12-2025
Pulished : 25-12-2025

Abstract

In the precision agriculture sector, the use of computer vision technology is a crucial solution to replace manual inspection methods. This research focuses on the problem of classifying Kecimen and Besni raisin varieties, which are often mixed up due to morphological similarities. The main objective of this study was to evaluate the effectiveness of the Logistic Regression algorithm in predicting raisin types and to analyze the impact of shape features on classification accuracy.The dataset used was the Raisin Dataset, which contains 900 data samples. Data preprocessing included feature normalization and attribute role assignment using RapidMiner data mining software. Seven morphological features extracted from digital images served as independent variables: area, perimeter, major and minor axis lengths, eccentricity, convexity area, and range. For model evaluation, a 70:30 split-data method was used, with 630 data samples used to train the model and 270 data samples prepared for performance testing.The experimental results showed that the Logistic Regression model achieved an overall accuracy of 84.07%. Further evaluation using a Confusion Matrix revealed balanced Precision and Recall values above 80% for both classes, indicating that the model exhibits no significant bias toward either variety. Although the misclassification rate was 15.93%, this is still acceptable considering the complexity of the biological similarities between varieties. These results support that Logistic Regression, as a computationally efficient linear method, is robust enough for use in real-time sorting systems compared to more complex and slow methods.

Keywords* : *Data Mining, Feature Extraction, Raisin Classification.

Abstrak

Dalam sektor pertanian presisi, penggunaan teknologi visi komputer menjadi solusi penting untuk menggantikan cara inspeksi manual. Penelitian ini fokus pada masalah pengelompokan varietas kismis Kecimen dan Besni yang sering tercampur karena kesamaan morfologi. Tujuan utama dari penelitian ini adalah untuk mengevaluasi seberapa efektif algoritma Logistic Regression dalam memprediksi jenis kismis serta menganalisis dampak fitur bentuk terhadap ketepatan klasifikasi. Dataset yang digunakan adalah Raisin Dataset yang berisi 900 contoh data. Proses pra-pemrosesan data mencakup normalisasi fitur dan penetapan peran atribut menggunakan perangkat lunak data mining RapidMiner. Tujuh fitur morfologis yang diambil dari gambar digital berfungsi sebagai variabel independen, yaitu luas area, keliling, panjang sumbu utama dan minor, eksentrisitas, luas cembung, dan jangkauan. Untuk evaluasi model, metode yang digunakan adalah Split Data dengan rasio 70:30, di mana 630 data dipakai untuk melatih model dan 270 data disiapkan untuk menguji kinerja. Hasil dari eksperimen menunjukkan bahwa model Logistic Regression mendapatkan akurasi keseluruhan sebesar 84,07%. Penilaian lebih lanjut dengan menggunakan Confusion Matrix mengungkapkan nilai Precision dan Recall yang seimbang di atas 80% untuk kedua kelas, yang menunjukkan bahwa model tidak menunjukkan bias yang signifikan terhadap salah satu dari varietas tersebut. Meskipun

terdapat tingkat kesalahan klasifikasi sebesar 15,93%, angka tersebut masih bisa diterima mengingat kerumitan kesamaan biologis antar varietas. Hasil ini mendukung bahwa Logistic Regression, sebagai metode linear yang efisien secara komputasi, cukup kuat untuk digunakan dalam sistem penyortiran real-time dibandingkan dengan metode yang lebih kompleks dan lambat.

Kata Kunci: Data Mining, Feature Extraction, Klasifikasi Kismis

PENDAHULUAN

Sektor pertanian modern kini semakin bergantung pada teknologi pasca-panen untuk menjamin standar kualitas produk dan efisiensi produksi. Kismis (*Vitis vinifera L.*) merupakan salah satu komoditas pertanian bernilai ekonomi tinggi yang diperdagangkan secara global. Sebagai salah satu produsen kismis terbesar di dunia, Turki memiliki berbagai varietas unggulan, di antaranya adalah varietas *Kecimen* dan *Besni*. Kedua varietas ini memiliki karakteristik rasa dan kegunaan yang berbeda, sehingga pemisahan (penyortiran) yang akurat sangat menentukan harga jual dan kepuasan konsumen di pasar internasional (Sadique et al., 2018) . Proses penjaminan mutu ini menuntut adanya sistem klasifikasi yang konsisten dan cepat.

Namun, permasalahan utama dalam industri pengolahan kismis adalah proses klasifikasi yang masih sering dilakukan secara manual atau menggunakan metode mekanis sederhana. Proses manual sangat bergantung pada persepsi visual manusia yang bersifat subjektif, rentan terhadap kelelahan, dan memakan waktu lama (*time-consuming*). Selain itu, varietas *Kecimen* dan *Besni* memiliki kemiripan morfologis yang sangat tinggi; keduanya memiliki warna dan tekstur yang hampir serupa, sehingga sulit dibedakan hanya dengan pengamatan sekilas oleh mata manusia. Kesalahan dalam penyortiran ini dapat menyebabkan tercampurnya varietas yang berujung pada penurunan nilai grade produk (ÇINAR et al., 2020). Oleh karena itu, diperlukan pendekatan berbasis teknologi cerdas yang mampu mengidentifikasi varietas kismis secara otomatis berdasarkan fitur-fitur fisik yang terukur.(Ayikpa et al., 2024)

Perkembangan pesat dalam bidang kecerdasan buatan (*Artificial Intelligence*) dan penambangan data (*Data Mining*) menawarkan solusi untuk mengatasi kendala tersebut. Teknik *Machine Vision* memungkinkan ekstraksi fitur morfologis—seperti luas area, keliling, keteraturan bentuk, dan panjang sumbu—menjadi data numerik yang dapat diolah (Sadique et al., 2018). Data fitur ini kemudian dapat dipelajari oleh algoritma *Machine Learning* untuk mengenali pola yang membedakan satu varietas dengan varietas lainnya. Pemanfaatan data mining dalam pertanian presisi (*precision agriculture*) telah terbukti mampu meningkatkan efisiensi operasional dan menekan biaya produksi secara signifikan.

Dalam konteks klasifikasi biner (dua kelas), metode *Logistic Regression* merupakan salah satu algoritma statistik yang paling mapan dan banyak digunakan. Berbeda dengan algoritma kompleks seperti *Neural Networks* yang seringkali menjadi "kotak hitam" (*black box*), *Logistic Regression* menawarkan keunggulan dalam hal interpretabilitas dan efisiensi komputasi. Algoritma ini bekerja dengan memodelkan probabilitas suatu entitas masuk ke dalam kelas tertentu berdasarkan hubungan linear antar variabel independennya. Karakteristik ini menjadikan *Logistic Regression* sangat cocok diterapkan pada kasus klasifikasi kismis, di mana tujuannya tidak hanya sekadar memprediksi label, tetapi juga memahami seberapa besar pengaruh dimensi fisik kismis terhadap jenis varietasnya.(Dreiseitl & Ohno-Machado, 2002)

Penelitian ini bertujuan untuk menerapkan dan mengevaluasi kinerja algoritma *Logistic Regression* dalam mengklasifikasikan varietas kismis *Kecimen* dan *Besni*. Penelitian ini memanfaatkan dataset publik dari *UCI Machine Learning Repository* yang berisi 900 sampel dengan 7 fitur morfologis hasil ekstraksi citra. Pengolahan data dilakukan menggunakan perangkat lunak RapidMiner Studio untuk memastikan alur kerja data mining yang sistematis, mulai dari pra-pemrosesan hingga evaluasi model. Hasil penelitian ini diharapkan dapat memberikan kontribusi empiris mengenai efektivitas metode statistik sederhana dalam menyelesaikan permasalahan klasifikasi pada produk pertanian, serta menjadi referensi bagi pengembangan sistem penyortiran otomatis yang hemat biaya.(ÇINAR et al., 2020)

METODE PENELITIAN

1. Alur Penelitian

Penelitian ini dilakukan dengan mengikuti kerangka kerja standar dalam penambangan data (*data mining*), yang terdiri dari lima tahapan utama: pengumpulan data, pra-pemrosesan data, pembagian data, pemodelan, dan evaluasi. Alur kerja penelitian secara skematis dapat dilihat pada uraian berikut:

a. Studi Literatur: Mempelajari teori terkait *Logistic Regression* dan karakteristik morfologi kismis.

b. Pengumpulan Data: Mengunduh dataset publik dari repositori terpercaya.

c. Pra-pemrosesan (*Preprocessing*): Menyiapkan data agar dapat diproses oleh algoritma, termasuk penetapan variabel target.

d. Pemodelan: Melatih algoritma *Logistic Regression* menggunakan data latih.

e. Evaluasi & Validasi: Menguji model menggunakan data uji dan mengukur performa menggunakan *Confusion Matrix*.(Ritthoff et al., 2001)

2. Data Penelitian

Data yang digunakan dalam penelitian ini adalah dataset sekunder "Raisin Dataset" yang diperoleh dari *UCI Machine Learning Repository*. Dataset ini merupakan hasil penelitian yang dipublikasikan oleh Cinar, Koklu, dan Tasdemir (2020).(ÇINAR et al., 2020)

Spesifikasi dataset adalah sebagai berikut:

Jumlah Sampel: 900 baris data (instances).

Jumlah Kelas: 2 kelas seimbang (*Balanced*), yaitu 450 sampel varietas *Kecimen* dan 450 sampel varietas *Besni*.

Atribut (Fitur): Terdiri dari 7 atribut numerik hasil ekstraksi fitur citra, yaitu:

a. *Area* (Luas wilayah kismis dalam piksel).

b. *Perimeter* (Keliling kismis).

c. *Major Axis Length* (Panjang garis sumbu utama).

d. *Minor Axis Length* (Panjang garis sumbu pendek).

JIIC: JURNAL INTELEK INSAN CENDIKIA
https://jicnusantara.com/index.php/jiic
Vol : 2 No: 12, Desember 2025
E-ISSN : 3047-7824

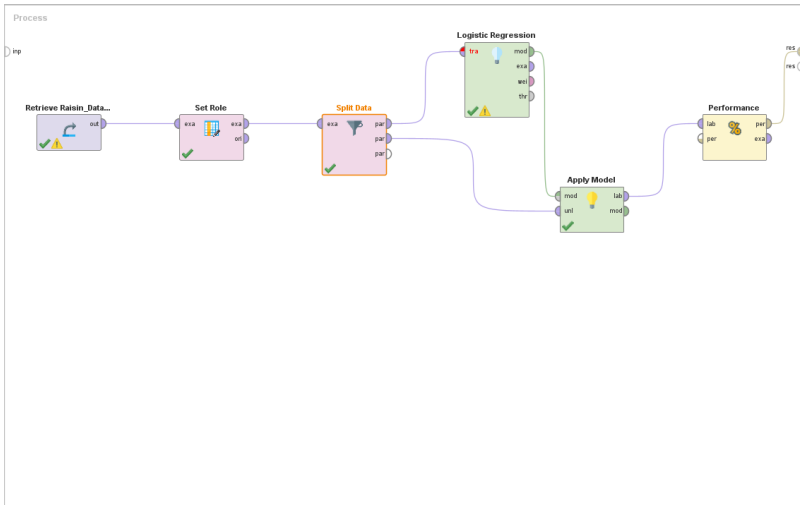
e. *Eccentricity* (Eksentrisitas/kelonjongan ellips).

f. *Convex Area* (Luas terkecil dari poligon cembung yang melingkupi kismis).

g. *Extent* (Rasio luas area terhadap *bounding box*).(Ritthoff et al., 2001)

3. Pra-pemrosesan Data (*Data Preprocessing*)

Tahap pra-pemrosesan dilakukan menggunakan perangkat lunak RapidMiner Studio. Langkah-langkah yang dilakukan meliputi:



(Sokolova & Lapalme, 2009)

a. Data Ingestion: Mengimpor file dataset (.csv atau .xlsx) ke dalam repositori RapidMiner.

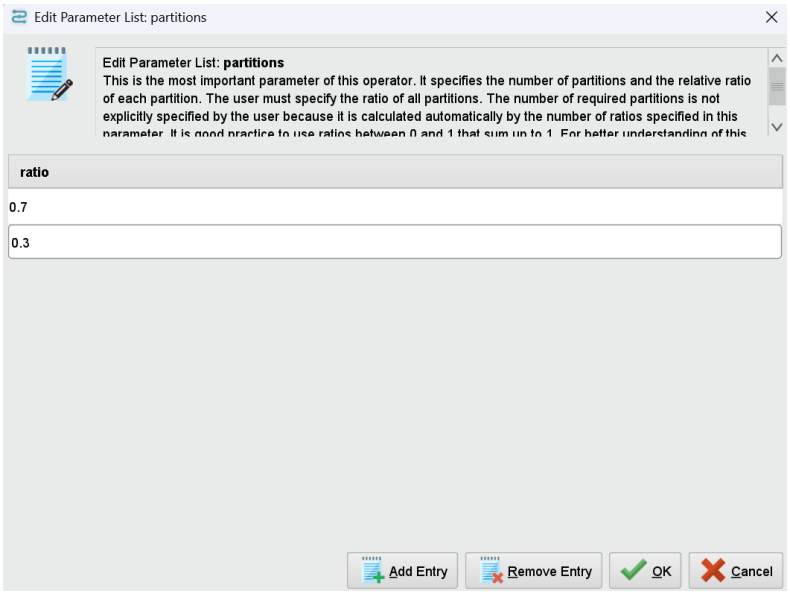
b. Set Role: Menggunakan operator *Set Role* untuk menetapkan atribut Class sebagai *Label* (variabel dependen/target) dan ketujuh atribut lainnya sebagai *Regular* (variabel independen/fitur).

c. Data Splitting (Pembagian Data): Untuk tujuan validasi, dataset dibagi menjadi dua bagian terpisah menggunakan operator *Split Data* dengan rasio pembagian 70:30.

1) Data Latih (70%): Sebanyak 630 data digunakan untuk melatih (*training*) model agar mempelajari pola hubungan antara fitur dan kelas.

JIIC: JURNAL INTELEK INSAN CENDIKIA
https://jicnusantara.com/index.php/jiic
Vol : 2 No: 12, Desember 2025
E-ISSN : 3047-7824

2) Data Uji (30%): Sebanyak 270 data disisihkan untuk menguji (*testing*) seberapa akurat model memprediksi data baru yang belum pernah dilihat sebelumnya.(Kohavi, 1995)



d. Metode Klasifikasi: Logistic Regression

Metode yang digunakan untuk membangun model klasifikasi adalah *Logistic Regression* (Regresi Logistik). Algoritma ini dipilih karena kemampuannya dalam menangani variabel dependen yang bersifat bier (dua kategori).(Dreiseitl & Ohno-Machado, 2002)

Model regresi logistik memprediksi probabilitas (P) bahwa sebuah instans data termasuk dalam kategori kelas "1" (misalnya: *Besni*) berdasarkan fungsi sigmoid. Persamaan fungsi logistik dinyatakan sebagai:

$$P(Y = 1)\frac{1}{1 + e^{-z}} \quad (1)$$

Dimana z adalah kombinasi linear dari fitur input (X) dan bobot (W) yang dipelajari model (z = w0 + w1x_1 + ... + w_nx_n). Jika nilai probabilitas P > 0.5, maka data akan diklasifikasikan ke dalam kelas target, dan sebaliknya.

e. Evaluasi Performa

Evaluasi model dilakukan dengan membandingkan hasil prediksi model terhadap label sebenarnya pada data uji. Instrumen evaluasi yang digunakan adalah **Confusion Matrix**, yang merepresentasikan empat kemungkinan hasil prediksi:

1) *True Positive (TP):* Data positif yang diprediksi benar.
2) *True Negative (TN):* Data negatif yang diprediksi benar.
3) *False Positive (FP):* Data negatif yang salah diprediksi sebagai positif.
4) *False Negative (FN):* Data positif yang salah diprediksi sebagai negatif.

Dari matriks tersebut, dihitung nilai Akurasi (*Accuracy*) sebagai metrik utama menggunakan rumus:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (2)$$

(Ritthoff et al., 2001)

f. Perangkat Penelitian

Perangkat keras dan lunak yang digunakan dalam penelitian ini adalah:

1) Perangkat Keras: Laptop dengan spesifikasi prosesor Intel Core i5/i7 (atau setara), RAM minimal 8GB, dan sistem operasi Windows 10/11.
2) Perangkat Lunak: RapidMiner Studio versi 9.x atau terbaru (versi Edukasi/Free) sebagai *platform* utama pengolahan data.

HASIL DAN PEMBAHASAN

Eksperimen klasifikasi varietas kismis dilakukan menggunakan algoritma *Logistic Regression* pada perangkat lunak RapidMiner. Dataset yang terdiri dari 900 sampel dibagi menjadi data latih sebanyak 630 sampel (70%) dan data uji sebanyak 270 sampel (30%). Proses pembelajaran model dilakukan pada data latih untuk mengenali pola fitur morfologis, kemudian model diuji validitasnya menggunakan data uji. Berdasarkan pengujian tersebut, diperoleh hasil performa model yang direpresentasikan dalam bentuk *Confusion Matrix* (Matriks Kebingungan). Matriks ini memetakan perbandingan antara prediksi kelas yang dihasilkan oleh model dengan kelas aktual (sebenarnya) dari data kismis. Ringkasan hasil prediksi pada 270 data uji dapat dilihat pada Tabel 1 di bawah ini.(Sokolova & Lapalme, 2009)

Table 1. *Confusion Matrix Hasil Klasifikasi Logistic Regression*

	True Kecimen	True Besni	Class Precision
Pred. Kecimen	112	20	84.85%
Pred. Besni	23	115	83.33%
Class Recall	82.96%	85.18%	Akurasi: 84.07%

Berdasarkan Tabel 1, model berhasil memprediksi dengan benar sebanyak 112 sampel kismis *Kecimen* dan 115 sampel kismis *Besni*. Total prediksi yang benar adalah 227 sampel dari keseluruhan 270 data uji. Sebaliknya, terjadi kesalahan prediksi (*misclassification*) pada 43 sampel, di mana 20 sampel *Besni* salah diprediksi sebagai *Kecimen*, dan 23 sampel *Kecimen* salah diprediksi sebagai *Besni*. Dari data tersebut, diperoleh nilai akurasi (*Accuracy*) keseluruhan sebesar 84,07%. Selain akurasi, metrik evaluasi lainnya menunjukkan nilai yang cukup seimbang, dengan rata-rata presisi (*Precision*) dan *Recall* berada di atas 82%.(Altuntaş et al., 2019)

Tingkat akurasi sebesar 84,07% mengindikasikan bahwa metode *Logistic Regression* memiliki kemampuan yang baik (*Good Classification*) dalam membedakan varietas kismis *Kecimen* dan *Besni* berdasarkan fitur fisiknya. Hal ini membuktikan bahwa atribut morfologis seperti luas area (*Area*), keliling (*Perimeter*), dan bentuk geometri (*Eccentricity*) memiliki korelasi yang signifikan terhadap jenis varietas kismis. Model mampu mempelajari bahwa varietas *Besni* cenderung memiliki bentuk yang lebih memanjang dibandingkan *Kecimen*, sehingga garis keputusan (*decision boundary*) linear yang dibentuk oleh fungsi logistik dapat memisahkan sebagian besar data dengan tepat.

Namun, terdapat tingkat kesalahan sebesar 15,93% yang tidak dapat diabaikan. Kesalahan klasifikasi ini terutama disebabkan oleh adanya irisan karakteristik (*feature overlapping*) antara kedua varietas. Dalam data biologis, variasi bentuk adalah hal yang alami; terdapat sebagian kecil kismis *Kecimen* yang berukuran lebih besar dari rata-rata sehingga menyerupai *Besni*, dan sebaliknya. Karena *Logistic Regression* bekerja dengan menarik garis pemisah lurus, algoritma ini kesulitan menangani data yang berada di area perbatasan yang ambigu tersebut. Meskipun demikian, nilai *Recall* untuk kelas *Besni* yang mencapai 85,18% menunjukkan bahwa model sedikit lebih sensitif dalam mengenali varietas *Besni* dibandingkan *Kecimen*.

Secara keseluruhan, hasil penelitian ini menegaskan bahwa penggunaan teknik data mining sederhana dapat menjadi alternatif solusi untuk efisiensi industri. Dibandingkan dengan penyortiran manual yang rentan subjektivitas, model ini menawarkan konsistensi penilaian dengan tingkat keberhasilan di atas 80%. Untuk meningkatkan akurasi di masa mendatang, disarankan untuk mengeksplorasi algoritma non-linear seperti *Support Vector Machine* (SVM) atau *Decision Tree* yang mungkin lebih mampu menangani data dengan batas kelas yang tidak terpisah secara linear.(Fawcett, 2006).

KESIMPULAN

Berdasarkan serangkaian percobaan dan analisis data yang telah dilakukan menggunakan metode *Logistic Regression* pada dataset kismis (*Raisin Dataset*), dapat ditarik beberapa kesimpulan penting. Pertama, implementasi algoritma *Logistic Regression* menggunakan perangkat lunak RapidMiner terbukti efektif untuk mengklasifikasikan varietas kismis *Kecimen* dan *Besni*. Dengan skema pembagian data latih 70% dan data uji 30%, model yang dibangun mampu menghasilkan tingkat akurasi sebesar **84,07%**. Angka ini menunjukkan bahwa fitur morfologis, seperti luas area dan dimensi sumbu, merupakan variabel prediktor yang kuat dalam membedakan identitas varietas kismis.

Kedua, analisis terhadap *Confusion Matrix* menunjukkan bahwa model memiliki performa yang relatif seimbang dalam memprediksi kedua kelas, dengan nilai *Recall* untuk kelas *Besni* (85,18%) sedikit lebih unggul dibandingkan kelas *Kecimen* (82,96%). Meskipun demikian, terdapat tingkat kesalahan klasifikasi sebesar 15,93%. Kesalahan ini diidentifikasi sebagai akibat dari adanya irisan karakteristik (*overlapping*) antara kedua varietas yang sulit dipisahkan secara linear sempurna. Hal ini wajar terjadi pada data biologis yang memiliki variabilitas alami, namun hasil di atas 80% tetap dikategorikan sebagai kinerja yang baik untuk sistem penyortiran otomatis awal.(Ritthoff et al., 2001)

Untuk pengembangan penelitian selanjutnya demi mencapai hasil yang lebih optimal, disarankan beberapa hal sebagai berikut. Pertama, peneliti selanjutnya dapat menerapkan metode validasi silang (*K-Fold Cross Validation*) untuk mendapatkan estimasi akurasi yang lebih stabil dan tidak bias terhadap satu kali pembagian data saja. Kedua, disarankan untuk membandingkan kinerja *Logistic Regression* dengan algoritma klasifikasi non-linear yang lebih kompleks, seperti *Support Vector Machine* (SVM) kernel RBF atau *Random Forest*, guna melihat peluang peningkatan akurasi pada data yang memiliki batas keputusan rumit. Terakhir, teknik seleksi fitur (*Feature Selection*) dapat diterapkan untuk mengeliminasi fitur yang kurang relevan, sehingga efisiensi komputasi model dapat ditingkatkan.(Wu et al., 2008)

DAFTAR PUSTAKA

Altuntaş, Y., Cömert, Z., & Kocamaz, A. F. (2019). Identification of haploid and diploid maize seeds using convolutional neural networks and a transfer learning approach. *Computers and Electronics in Agriculture*, *163*, 104874. https://doi.org/10.1016/J.COMPAG.2019.104874

Ayikpa, K. J., Gouton, P., Mamadou, D., & Ballo, A. B. (2024). Classification of Cocoa Beans by Analyzing Spectral Measurements Using Machine Learning and Genetic Algorithm. *Journal of Imaging*, *10*(1). https://doi.org/10.3390/jimaging10010019

ÇINAR, İ., KOKLU, M., & TAŞDEMİR, Ş. (2020). Kuru Üzüm Tanelerinin Makine Görüşü ve Yapay Zeka Yöntemleri Kullanılarak Sınıflandırılması. *Gazi Journal of Engineering Sciences*, *6*(3), 200–209. https://doi.org/10.30855/gmbd.2020.03.03

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics*, *35*(5–6), 352–359. https://doi.org/10.1016/S1532-0464(03)00034-0

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/J.PATREC.2005.10.010

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *IJCAI International Joint Conference on Artificial Intelligence*, *2*, 1137–1143.

Ritthoff, O., Klinkenberg, R., Fischer, S., Mierswa, I., & Felske, S. (2001). Yale: Yet Another Machine Learning Environment. *LLWA 01 -- Tagungsband Der GI-Workshop-Woche Lernen -- Lehren -- Wissen -- Adaptivit□t Number Nr. 763 in Series Forschungsberichte Des Fachbereichs Informatik, Universit□t Dortmund*, 84–92.

Sadique, K. M., Rahmani, R., & Johannesson, P. (2018). Towards Security on Internet of Things: Applications and Challenges in Technology. *Procedia Computer Science*, *141*, 199–206. https://doi.org/10.1016/J.PROCS.2018.10.168

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. https://doi.org/10.1016/J.IPM.2009.03.002

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., & Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1–37. https://doi.org/10.1007/s10115-007-0114-2