Analisis Algoritma Logistic Regression dan Support Vector Machine pada Kasus Klasifikasi Citra Hewan Rawa dengan Dataset yang tidak Seimbang

Dedy Armiady[1)*]
[1)] Program Studi Sistem Informasi, Universitas Almuslim Bireuen

Received: 1 Aug 2024 Accepted: 05 Aug 2024 Published: 13 Aug 2024

***dedy.armiady@gmail.com**

Kata Kunci: Klasifikasi Citra, Logistic Regression, Support Vector Machine, Ketidakseimbangan Data.

Abstrak : Penelitian ini bertujuan untuk mengevaluasi kinerja dua algoritma machine learning, Logistic Regression dan Support Vector Machine (SVM), dalam tugas klasifikasi citra hewan rawa menggunakan dataset yang tidak seimbang. Dataset yang digunakan terdiri dari empat kategori hewan rawa: Buaya, Kodok, Kura-kura, dan Ular, dengan distribusi yang sangat tidak merata. Kelas Kura-kura memiliki jumlah sampel yang jauh lebih banyak dibandingkan dengan kelas lainnya, menciptakan tantangan ketidakseimbangan data yang signifikan. Metode penelitian dimulai dengan mengimpor dataset citra hewan rawa ke dalam tool Orange Data Mining, diikuti oleh proses ekstraksi fitur menggunakan SqueezeNet (local) sebagai embedder. Dua model machine learning, yaitu Logistic Regression dan SVM, kemudian dilatih menggunakan fitur yang diekstraksi. Evaluasi model dilakukan dengan menambahkan widget test and score untuk mengukur metrik performa seperti Area Under the Curve (AUC), akurasi klasifikasi (CA), F1-Score, precision, recall, dan Matthews Correlation Coefficient (MCC). Hasil penelitian menunjukkan bahwa Logistic Regression unggul dalam hampir semua metrik evaluasi dibandingkan SVM. Logistic Regression mencapai nilai AUC sebesar 0.985, akurasi klasifikasi 0.908, F1-Score 0.909, precision 0.909, recall 0.908, dan MCC 0.859. Sebaliknya, SVM mencapai nilai AUC 0.971, akurasi klasifikasi 0.863, F1-Score 0.867, precision 0.877, recall 0.863, dan MCC 0.797. Kesimpulan dari penelitian ini adalah bahwa Logistic Regression merupakan model yang lebih tepat untuk tugas klasifikasi citra hewan rawa dengan dataset yang tidak seimbang. Model ini tidak hanya menunjukkan kinerja yang lebih baik dalam membedakan kelas-kelas citra tetapi juga lebih akurat dan seimbang dalam mengklasifikasikan sampel dari kelas minoritas.

PENDAHULUAN

Penggunaan machine learning dalam klasifikasi citra telah menunjukkan potensi besar dalam berbagai aplikasi, termasuk pengenalan hewan [1]. Dalam hal ini, klasifikasi citra hewan rawa menjadi salah satu tantangan menarik yang perlu dieksplorasi lebih lanjut. Hewan-hewan rawa seperti buaya, kodok, kura-kura, dan ular memiliki karakteristik visual yang berbeda-beda, sehingga memerlukan pendekatan yang tepat untuk mengidentifikasi setiap spesies secara akurat.

Salah satu tantangan utama dalam penelitian ini adalah ketidakseimbangan data. Dataset yang digunakan dalam penelitian ini berasal dari Kaggle dan terdiri dari empat kelas hewan rawa dengan jumlah yang bervariasi, yaitu Buaya (692), Kodok (497), Kura-kura (1862), dan Ular (500). Ketidakseimbangan ini dapat mempengaruhi kinerja algoritma klasifikasi, karena model cenderung lebih mudah belajar dari kelas dengan jumlah data yang lebih banyak dibandingkan kelas dengan data yang lebih sedikit.

Dalam penelitian ini, dua algoritma machine learning, yaitu Logistic Regression dan Support Vector Machine (SVM), akan dianalisis untuk melihat kinerjanya dalam klasifikasi citra hewan rawa. Logistic Regression dikenal sebagai algoritma yang sederhana namun efektif untuk tugas klasifikasi, sedangkan SVM

terkenal dengan kemampuannya menangani masalah klasifikasi dengan margin yang jelas antara kelas-kelas. Kedua algoritma ini memiliki karakteristik dan pendekatan yang berbeda dalam memproses data, sehingga menarik untuk dibandingkan kinerjanya.

Adapun tujuan dari penelitian ini adalah untuk mengevaluasi bagaimana kedua algoritma tersebut menangani ketidakseimbangan data dalam tugas klasifikasi citra hewan rawa. Dengan mengukur kinerja kedua algoritma ini, penelitian ini diharapkan dapat memberikan wawasan tentang kelebihan dan kekurangan masing-masing metode pada kasus dataset yang tidak merata. Hasil penelitian ini akan berguna bagi pengembangan lebih lanjut dalam bidang klasifikasi citra dan aplikasi machine learning lainnya yang menghadapi tantangan serupa. Selain itu, penelitian ini juga berupaya untuk memberikan rekomendasi tentang algoritma yang lebih cocok digunakan dalam situasi di mana data training tidak merata. Dengan demikian, hasil penelitian ini tidak hanya cocok untuk klasifikasi citra hewan rawa, tetapi juga dapat diterapkan dalam berbagai kasus klasifikasi citra lainnya yang menghadapi masalah ketidakseimbangan data..

TINJAUAN LITERATUR

Klasifikasi citra merupakan salah satu bidang utama dalam penelitian machine learning yang memiliki berbagai aplikasi, termasuk pengenalan hewan. Algoritma seperti Logistic Regression dan Support Vector Machine (SVM) telah banyak digunakan dalam tugas-tugas klasifikasi karena kemampuan mereka dalam menangani data yang bervariasi dan kompleks. Logistic Regression dikenal karena kesederhanaan dan efisiensinya dalam tugas-tugas klasifikasi linier [2], sementara SVM terkenal dengan kemampuannya untuk memisahkan data dengan margin yang maksimal, bahkan dalam ruang dimensi tinggi [3]. Salah satu tantangan utama dalam klasifikasi citra adalah ketidakseimbangan data, di mana jumlah data dalam setiap kelas tidak merata. Ketidakseimbangan ini dapat mengakibatkan bias pada model yang lebih cenderung memprediksi kelas dengan data lebih banyak [4]. Berbagai teknik telah dikembangkan untuk mengatasi masalah ini, termasuk penyeimbangan data melalui resampling, penyesuaian bobot pada algoritma, dan penggunaan metrik evaluasi yang lebih tepat seperti F1-score dan area under the ROC curve (AUC) [5], [6], [7]. Dalam masalah klasifikasi citra hewan, penelitian oleh [8] menunjukkan bahwa SVM dapat memberikan hasil yang lebih baik dibandingkan dengan Logistic Regression, terutama ketika dataset memiliki fitur yang tidak linier. Studi ini menekankan pentingnya pemilihan kernel yang tepat dalam SVM untuk menangani kompleksitas data citra. Di sisi lain, Logistic Regression tetap menjadi pilihan yang populer karena interpretabilitasnya dan kemampuannya untuk menangani masalah klasifikasi yang sederhana dan linier [9].

Penelitian tentang klasifikasi citra hewan rawa dengan dataset yang tidak seimbang masih relatif terbatas. Namun, penelitian oleh [10] mengindikasikan bahwa pendekatan hybrid yang menggabungkan metode penyeimbangan data dengan algoritma klasifikasi dapat meningkatkan kinerja model. Studi ini menemukan bahwa kombinasi teknik oversampling dengan SVM menghasilkan peningkatan signifikan dalam akurasi klasifikasi pada dataset yang tidak seimbang. Penelitian oleh [11] menunjukkan bahwa menggunakan ensemble learning dapat mengurangi efek ketidakseimbangan data dan meningkatkan akurasi klasifikasi. Namun, metode ini juga memiliki kelemahan, terutama dalam hal interpretabilitas dan kompleksitas komputasi. Dalam studi oleh [12], mereka mengevaluasi berbagai pendekatan untuk menangani ketidakseimbangan data dalam klasifikasi citra, termasuk penggunaan algoritma boosting dan bagging. Mereka menemukan bahwa metode boosting, seperti AdaBoost, dapat membantu memperbaiki performa model pada kelas minoritas tanpa mengorbankan akurasi pada kelas mayoritas.

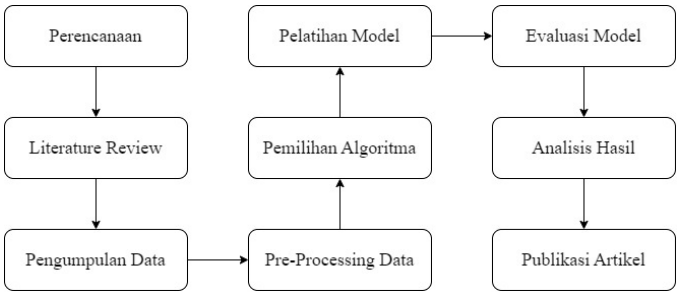
Dalam penelitian ini, kami akan mengevaluasi kinerja Logistic Regression dan SVM dalam klasifikasi citra hewan rawa dengan menggunakan dataset dari Kaggle yang terdiri dari empat kelas: Buaya, Kodok, Kura-kura, dan Ular. Kami akan mengukur kinerja kedua algoritma tersebut dengan mempertimbangkan metrik evaluasi yang sesuai untuk dataset yang tidak seimbang, seperti precision, recall, dan F1-score [13]. Penelitian lain oleh [14] juga menunjukkan bahwa menggunakan metode oversampling seperti ADASYN (Adaptive Synthetic Sampling Approach for Imbalanced Learning) dapat meningkatkan kinerja model pada dataset yang tidak seimbang. Teknik ini menyesuaikan distribusi data sintetis berdasarkan kesulitan klasifikasi sampel yang ada, yang dapat membantu meningkatkan generalisasi model. Untuk lebih memahami dampak ketidakseimbangan data, studi oleh [15] melakukan analisis sistematis terhadap berbagai teknik penyeimbangan data dan menemukan bahwa tidak ada satu metode yang cocok untuk semua situasi. Oleh

karena itu, penting untuk mengevaluasi beberapa pendekatan dan menyesuaikannya dengan karakteristik spesifik dari dataset yang digunakan.

Penelitian lain oleh [16] mengusulkan penggunaan algoritma k-Nearest Neighbors (k-NN) dalam kombinasi dengan teknik oversampling dan undersampling untuk menangani ketidakseimbangan data. Mereka menemukan bahwa metode ini dapat membantu meningkatkan akurasi klasifikasi pada kelas minoritas tanpa mengorbankan performa pada kelas mayoritas. Sebagai tambahan, penelitian oleh [17] mengeksplorasi penggunaan teknik ensemble yang menggabungkan beberapa model klasifikasi untuk menangani ketidakseimbangan data. Mereka menemukan bahwa metode ensemble dapat membantu meningkatkan robustnes dan generalisasi model, terutama dalam kasus dengan ketidakseimbangan data yang parah. Dalam studi oleh [18], mereka mengevaluasi kinerja berbagai algoritma klasifikasi pada dataset yang tidak seimbang dan menemukan bahwa algoritma seperti Decision Trees dan Neural Networks juga dapat dipertimbangkan sebagai alternatif untuk Logistic Regression dan SVM. Mereka menemukan bahwa dengan penyesuaian parameter yang tepat, algoritma ini dapat memberikan hasil yang kompetitif. Penelitian oleh [19] juga menunjukkan bahwa penggunaan teknik augmentasi data dapat membantu meningkatkan kinerja model pada dataset yang tidak seimbang. Mereka menemukan bahwa dengan menambahkan variasi pada data pelatihan, model dapat belajar representasi yang lebih baik dan meningkatkan akurasi klasifikasi. Terakhir, penelitian oleh [20] menyarankan bahwa pemilihan metrik evaluasi yang tepat sangat penting dalam mengevaluasi kinerja model pada dataset yang tidak seimbang. Mereka merekomendasikan penggunaan metrik seperti precision, recall, F1-score, dan AUC untuk mendapatkan gambaran yang lebih akurat tentang kinerja model.

METODE PENELITIAN

Penelitian ini bertujuan untuk menganalisis kinerja algoritma Logistic Regression dan Support Vector Machine (SVM) dalam klasifikasi citra hewan rawa menggunakan dataset yang tidak seimbang. Metode penelitian yang digunakan terdiri dari beberapa tahap utama sebagai berikut:



Gambar 1 Alur Penelitian

Penelitian ini dimulai dari tahap perencanaan, di mana pada tahap ini tujuan dan lingkup penelitian ditentukan. Identifikasi masalah penelitian, sasaran, dan metodologi yang akan digunakan dilakukan untuk memastikan arah yang jelas dan terstruktur. Selanjutnya, penetapan tujuan penelitian dilakukan berdasarkan masalah yang telah diidentifikasi, untuk menetapkan sasaran spesifik yang ingin dicapai melalui penelitian ini. Tahap berikutnya adalah pemilihan dataset. Data yang akan digunakan dalam penelitian ini dikumpulkan dari sumber sekunder, yaitu dataset publik yang tersedia di Kaggle. Dataset ini terdiri dari citra hewan rawa dengan empat kelas: Buaya, Kodok, Kura-kura, dan Ular. Setelah dataset terkumpul, tahap penentuan algoritma dan tools dilakukan dengan memilih algoritma machine learning yang sesuai dan alat yang akan digunakan untuk analisis data, dalam hal ini, Logistic Regression dan Support Vector Machine (SVM).

Pada tahap pre-processing data, data yang telah dikumpulkan dibersihkan dan dipersiapkan agar siap digunakan dalam pemodelan. Proses ini mencakup penanganan data yang hilang, normalisasi data untuk memastikan skala fitur yang konsisten, serta transformasi data untuk meningkatkan kualitas informasi. Setelah data siap, tahap pengembangan model dimulai dengan mengembangkan model machine learning berdasarkan algoritma yang dipilih.

Tahap pelatihan model dilakukan dengan menggunakan dataset yang telah diproses untuk melatih model machine learning. Dataset dibagi menjadi data pelatihan dan data pengujian untuk memastikan model dapat diuji dengan data yang tidak digunakan selama pelatihan. Validasi model dilakukan dengan menggunakan teknik validasi silang (cross-validation) untuk memastikan model dapat bekerja dengan baik pada data baru dan menghindari overfitting. Keseluruhan proses pelatihan menggunakan tool Orange Data Mining untuk fleksibilitas fitur dan kemudahan training. Evaluasi performa model dilakukan dengan menggunakan metrik yang sesuai seperti precision, recall, F1-score, dan area under the ROC curve (AUC). Evaluasi ini penting untuk memberikan gambaran tentang seberapa baik model bekerja pada data yang tidak seimbang. Hasil evaluasi kemudian dianalisis untuk menarik kesimpulan tentang kinerja algoritma yang digunakan, yaitu Logistic Regression dan SVM.

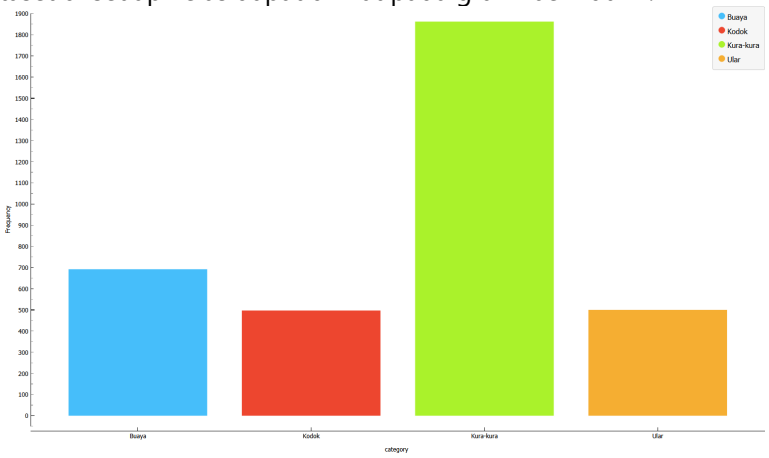
Setelah analisis hasil, tahap penyusunan laporan penelitian dilakukan. Laporan ini mencakup latar belakang, metodologi, hasil, dan diskusi, serta semua temuan penting dan kesimpulan dari penelitian. Laporan yang disusun kemudian digunakan sebagai dasar untuk menulis artikel penelitian yang akan dikirimkan ke jurnal ilmiah untuk dipublikasikan. Tahap akhir penelitian ini melibatkan proses review oleh rekan sejawat dan revisi berdasarkan umpan balik yang diterima untuk memastikan kualitas dan kredibilitas publikasi ilmiah.

Dataset yang digunakan dalam penelitian ini adalah dataset gambar hewan rawa yang dikumpulkan dari sumber terbuka Kaggle, dan tersebar dalam 4 kelas. Adapun rincian dataset yang digunakan dapat dilihat pada tabel berikut:

Tabel 1. Dataset

Jenis	Format Data	Kelas	Jumlah Data
Data Training	JPG	Buaya	692
		Kodok	497
		Kura-kura	1862
		Ular	500
Total Dataset			3551

Adapun sebaran dataset di setiap kelas dapat dilihat pada grafik berikut ini:

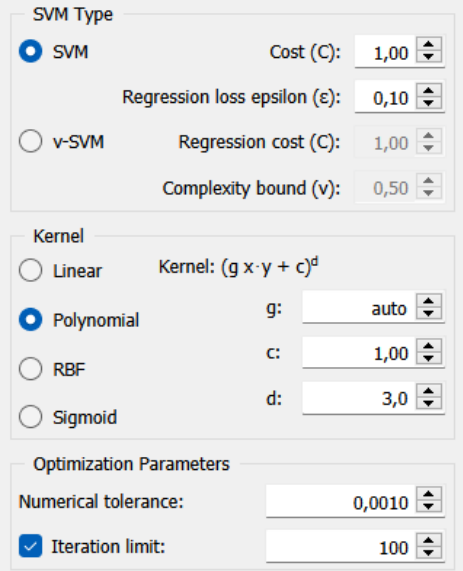


Gambar 2. Sebaran Dataset

Gambar 2 adalah gambar diagram batang yang menunjukkan distribusi frekuensi dari empat kategori hewan rawa dalam dataset. Kategori buaya memiliki frekuensi sekitar 692 dengan warna biru, menunjukkan jumlah data yang moderat dibandingkan dengan kategori lainnya. Kodok, yang diwakili oleh warna merah, memiliki frekuensi sekitar 497 dan merupakan kategori dengan jumlah data paling sedikit di antara semua kategori. Ini menunjukkan bahwa kodok adalah kelas minoritas dalam dataset ini, yang dapat menyebabkan tantangan dalam pelatihan model machine learning karena model mungkin akan kesulitan mempelajari representasi yang baik untuk kelas ini. Kategori kura-kura, diwakili oleh warna hijau, memiliki frekuensi sekitar 1862, menjadikannya kelas mayoritas dalam dataset. Jumlah data yang besar untuk kelas ini memungkinkan

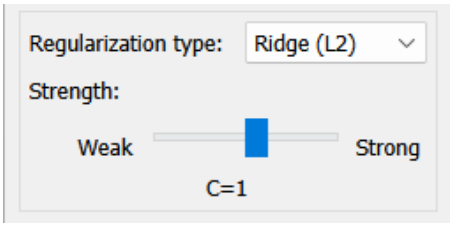
model machine learning belajar lebih baik, namun dapat menyebabkan model bias terhadap kelas mayoritas. Kategori ular, dengan frekuensi sekitar 500 dan warna oranye, juga termasuk kelas minoritas, mirip dengan kategori kodok. Adapun parameter daripada algoritma yang digunakan adalah sebagai berikut:

1) Parameter Support Vector Machine (SVM):



Gambar 3. Parameter SVM

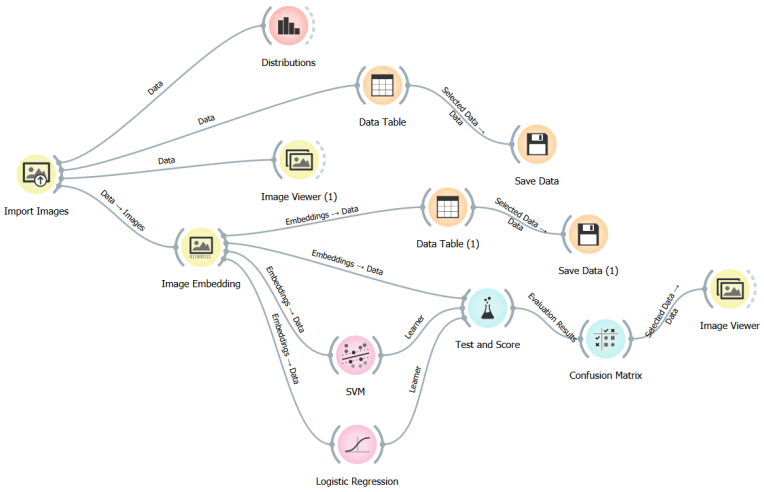
2) Parameter Logistic Regression:



Gambar 4. Parameter Logistic Regression

HASIL PENELITIAN

Pada penelitian ini, tool Orange Data Mining digunakan untuk melakukan klasifikasi citra hewan rawa dengan memanfaatkan dua model machine learning, yaitu Logistic Regression dan Support Vector Machine (SVM). Proses penelitian dimulai dengan mengimpor dataset citra hewan rawa menggunakan widget import image, dilanjutkan dengan penambahan widget image embedding menggunakan SqueezeNet (local) untuk mengekstrak fitur dari citra. Dua model machine learning, Logistic Regression dan SVM, kemudian dilatih menggunakan fitur yang diekstraksi. Evaluasi performa model dilakukan menggunakan widget test and score, menghasilkan perbedaan akurasi dan metrik evaluasi lainnya antara kedua model. Berikut adalah gambar dari lingkungan kerja tool Orange Data Mining untuk memproses training dengan kedua algoritma yang digunakan dalam penelitian ini:



Gambar 5. Proses Klasifikasi

Adapun hasil klasifikasi dari kedua algoritma menggunakan parameter yang telah disebutkan sebelumnya, dapat dilihat pada gambar berikut:

Model	AUC	CA	F1	Prec	Recall	MCC
SVM	0.971	0.863	0.867	0.877	0.863	0.797
Logistic Regression	0.985	0.908	0.909	0.909	0.908	0.859

Gambar 6. Hasil Klasifikasi

Berdasarkan hasil pengujian yang ditampilkan pada gambar, analisis menunjukkan bahwa Logistic Regression unggul dalam beberapa metrik evaluasi dibandingkan SVM. Pada metrik Area Under the Curve (AUC), Logistic Regression memperoleh nilai 0.985, lebih tinggi dibandingkan SVM yang memperoleh nilai 0.971. Ini menunjukkan bahwa Logistic Regression memiliki kemampuan yang sedikit lebih baik dalam membedakan antara kelas-kelas dalam dataset. Akurasi klasifikasi (CA) juga menunjukkan keunggulan Logistic Regression dengan nilai 0.908 dibandingkan SVM yang memiliki nilai 0.863, mengindikasikan bahwa Logistic Regression lebih akurat dalam mengklasifikasikan citra hewan rawa. F1-Score, yang merupakan metrik harmonis dari precision dan recall, menunjukkan bahwa Logistic Regression memiliki nilai 0.909, lebih tinggi dibandingkan SVM yang memiliki nilai 0.867. Hal ini menunjukkan bahwa Logistic Regression memiliki keseimbangan yang lebih baik antara precision dan recall. Precision sendiri, yang menunjukkan proporsi prediksi positif yang benar dari semua prediksi positif, lebih tinggi pada Logistic Regression dengan nilai 0.909 dibandingkan SVM yang memiliki nilai 0.877, menunjukkan bahwa Logistic Regression lebih baik dalam mengurangi jumlah false positives. Pada metrik recall, yang menunjukkan proporsi prediksi positif yang benar dari semua sampel positif, Logistic Regression juga unggul dengan nilai 0.908 dibandingkan SVM yang memiliki nilai 0.863. Ini menunjukkan bahwa Logistic Regression lebih baik dalam menangkap semua sampel positif. Matthews Correlation Coefficient (MCC), yang memperhitungkan true positives, false negatives, true negatives, dan false positives, juga lebih tinggi pada Logistic Regression dengan nilai 0.859 dibandingkan SVM yang memiliki nilai 0.797, menunjukkan performa keseluruhan yang lebih baik.

PEMBAHASAN

Dalam penelitian ini, dataset yang digunakan untuk klasifikasi citra hewan rawa menunjukkan ketidakseimbangan yang signifikan antara kelas-kelas yang berbeda. Dataset terdiri dari empat kategori, yaitu Buaya, Kodok, Kura-kura, dan Ular. Kelas Kura-kura memiliki jumlah data yang jauh lebih banyak (1862 sampel) dibandingkan dengan kelas lainnya seperti Buaya (692 sampel), Kodok (497 sampel), dan Ular (500 sampel). Ketidakseimbangan data ini dapat menimbulkan beberapa tantangan dalam pelatihan model machine learning. Ketidakseimbangan data dapat menyebabkan model machine learning menjadi bias

terhadap kelas mayoritas, dalam hal ini Kura-kura, karena model lebih banyak belajar dari kelas yang memiliki lebih banyak sampel. Akibatnya, model mungkin tidak dapat mempelajari representasi yang baik untuk kelas minoritas, seperti Kodok dan Ular, sehingga menghasilkan performa yang buruk dalam mengklasifikasikan sampel dari kelas-kelas minoritas tersebut. Oleh karena itu, penting untuk memilih dan mengatur model dengan baik agar dapat menangani ketidakseimbangan data ini.

Berdasarkan hasil evaluasi, Logistic Regression menunjukkan performa yang lebih baik dibandingkan dengan SVM dalam hampir semua metrik evaluasi yang digunakan. Berikut adalah beberapa alasan mengapa Logistic Regression lebih baik daripada SVM dalam hal dataset yang tidak seimbang ini:

1) Penanganan Ketidakseimbangan Data
   Logistic Regression secara alami dapat menangani ketidakseimbangan data dengan lebih baik karena model ini menggunakan probabilitas untuk mengklasifikasikan sampel. Dengan menyesuaikan ambang batas keputusan, Logistic Regression dapat meningkatkan recall pada kelas minoritas tanpa mengorbankan precision secara signifikan. Hal ini membantu model untuk lebih responsif terhadap sampel dari kelas minoritas.
2) Simplicity and Interpretability
   Logistic Regression adalah model yang relatif sederhana dan mudah diinterpretasikan. Model ini tidak memerlukan banyak penyesuaian parameter dan biasanya bekerja dengan baik pada dataset yang tidak terlalu kompleks. Sementara itu, SVM, terutama dengan kernel yang kompleks, memerlukan penyesuaian parameter yang lebih teliti (seperti pemilihan kernel dan penyesuaian parameter regularisasi), yang bisa menjadi sulit dan tidak selalu menghasilkan performa yang lebih baik pada dataset yang tidak seimbang.
3) Overfitting
   Logistic Regression cenderung lebih tahan terhadap overfitting dibandingkan SVM pada dataset yang tidak seimbang. SVM dengan kernel yang kompleks dapat overfit terhadap kelas mayoritas, terutama ketika kelas mayoritas memiliki variasi yang lebih banyak dalam data. Overfitting ini menyebabkan SVM menjadi kurang generalizable pada kelas minoritas, sehingga menghasilkan performa yang lebih buruk pada metrik seperti recall dan F1-Score.
4) Flexibility in Threshold Adjustment
   Logistic Regression memungkinkan penyesuaian ambang batas keputusan untuk mengoptimalkan trade-off antara precision dan recall. Ini sangat berguna pada kasus dataset yang tidak seimbang di mana prioritas mungkin diberikan pada recall untuk memastikan bahwa sebanyak mungkin sampel dari kelas minoritas dapat diidentifikasi dengan benar. Sementara itu, SVM tidak memiliki fleksibilitas yang sama dalam penyesuaian ambang batas keputusan, yang dapat membatasi kemampuannya untuk menyesuaikan performa pada dataset yang tidak seimbang.
5) Evaluasi dan Metrik
   Dari hasil evaluasi yang ditunjukkan, Logistic Regression memiliki AUC, akurasi, F1-Score, precision, recall, dan MCC yang lebih tinggi dibandingkan dengan SVM. Ini menunjukkan bahwa Logistic Regression tidak hanya lebih baik dalam membedakan antara kelas-kelas (AUC) tetapi juga lebih akurat secara keseluruhan dan lebih seimbang dalam penanganan precision dan recall.

Logistic Regression dan Support Vector Machine (SVM) adalah dua algoritma machine learning yang sering digunakan untuk tugas klasifikasi. Meskipun mereka berbeda dalam pendekatan teknis dan penerapan, kedua algoritma ini memiliki beberapa persamaan. Pertama, baik Logistic Regression maupun SVM digunakan untuk memprediksi kelas dari data yang diberikan berdasarkan fitur-fitur yang ada. Keduanya adalah algoritma supervised learning, yang berarti mereka memerlukan data yang sudah dilabeli untuk melatih model. Kedua, kedua algoritma ini dapat menangani data yang linier dan dapat diperluas untuk menangani data non-linier melalui penggunaan teknik seperti basis fungsi kernel dalam SVM dan regularisasi dalam Logistic Regression. Selanjutnya, kedua algoritma ini menghasilkan output yang dapat digunakan untuk mengukur probabilitas atau margin keputusan, yang membantu dalam menentukan kelas dari sampel yang diuji. Logistic Regression mengeluarkan probabilitas langsung dari kelas positif, sedangkan SVM mengeluarkan nilai margin keputusan yang dapat diubah menjadi probabilitas melalui metode seperti Platt scaling. Selain itu, baik Logistic Regression maupun SVM memiliki mekanisme untuk menangani

masalah overfitting. Logistic Regression menggunakan teknik regularisasi seperti L1 dan L2, sementara SVM mengatur kompleksitas model melalui parameter regularisasi C dan pemilihan kernel yang tepat. Kedua algoritma ini juga mendukung berbagai metrik evaluasi untuk mengukur kinerja model, seperti precision, recall, F1-Score, dan area under the curve (AUC). Mereka dapat diterapkan pada berbagai jenis data dan masalah klasifikasi, mulai dari pengenalan citra hingga analisis teks. Meskipun pendekatan matematis dan teknik optimisasi mereka berbeda, Logistic Regression dan SVM sama-sama bertujuan untuk memaksimalkan akurasi prediksi dengan meminimalkan kesalahan klasifikasi, sehingga membuat mereka menjadi pilihan populer dalam berbagai aplikasi machine learning.

KESIMPULAN

Penelitian ini menunjukkan bahwa Logistic Regression adalah model yang lebih unggul dibandingkan Support Vector Machine (SVM) dalam tugas klasifikasi citra hewan rawa dengan dataset yang tidak seimbang. Dalam proses penelitian yang menggunakan tool Orange Data Mining, model Logistic Regression berhasil mencapai nilai metrik evaluasi yang lebih tinggi dibandingkan SVM. Logistic Regression menunjukkan performa yang lebih baik dalam membedakan kelas-kelas citra, dengan nilai Area Under the Curve (AUC) yang lebih tinggi. Selain itu, model ini juga mencapai akurasi klasifikasi yang lebih tinggi, serta F1-Score, precision, recall, dan Matthews Correlation Coefficient (MCC) yang lebih baik. Ketidakseimbangan data dalam dataset ini, di mana kelas mayoritas (Kura-kura) mendominasi jumlah sampel, dapat menyebabkan model machine learning menjadi bias terhadap kelas mayoritas. Logistic Regression, dengan kemampuannya yang lebih baik dalam menangani ketidakseimbangan data dan mengoptimalkan trade-off antara precision dan recall, terbukti lebih efektif dibandingkan SVM. Oleh karena itu, berdasarkan hasil penelitian ini, Logistic Regression dapat dianggap sebagai model yang lebih tepat untuk tugas klasifikasi citra hewan rawa, memberikan hasil yang lebih akurat dan seimbang dalam menghadapi tantangan ketidakseimbangan data.

REFERENCES

[1] L. Luthfi, R. Imam Muslem, D. Armiady, S. Sriwinar, R. Fajri, and I. Iqbal, "Analysis of CNN Method for Image Classification of Coconut Ripeness Levels," in *2023 Eighth International Conference on Informatics and Computing (ICIC)*, IEEE, Dec. 2023, pp. 1–6. doi: 10.1109/ICIC60109.2023.10381964.

[2] S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," 2007.

[3] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Min Knowl Discov*, vol. 2, no. 2, 1998, doi: 10.1023/A:1009715923555.

[4] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans Knowl Data Eng*, vol. 21, no. 9, 2009, doi: 10.1109/TKDE.2008.239.

[5] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," 2016. doi: 10.1007/s13748-016-0094-0.

[6] D. Armiady and I. M. R, "Klasifikasi Kualitas Buah Pisang Berdasarkan Citra Buah Menggunakan Stochastic Gradient Descent," *KLIK: Kajian Ilmiah Informatika dan Komputer*, vol. 4, no. 2, 2023.

[7] I. R. Muslem, "KLIK: Kajian Ilmiah Informatika dan Komputer Image Classification pada Kasus American Sign Language Menggunakan Support Vector Machine," *Media Online)*, vol. 4, no. 2, pp. 1184–1191, 2023, doi: 10.30865/klik.v4i2.1242.

[8] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artif Intell Rev*, vol. 53, no. 8, 2020, doi: 10.1007/s10462-020-09825-6.

[9] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: A statistical view of boosting," 2000. doi: 10.1214/aos/1016218223.

[10] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proceedings of the International Joint Conference on Neural Networks*, 2008. doi: 10.1109/IJCNN.2008.4633969.

[11] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," 2012. doi: 10.1109/TSMCC.2011.2161285.

[12] L. Liu, P. Wang, J. Lin, and L. Liu, "Intrusion Detection of Imbalanced Network Traffic Based on Machine Learning and Deep Learning," *IEEE Access*, vol. 9, 2021, doi: 10.1109/ACCESS.2020.3048198.

[13] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf Process Manag*, vol. 45, no. 4, 2009, doi: 10.1016/j.ipm.2009.03.002.

[14] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data:

Review of methods and applications," 2017. doi: 10.1016/j.eswa.2016.12.035.
[15] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, vol. 6, no. 5, 2002, doi: 10.3233/ida-2002-6504.
[16] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, 2004, doi: 10.1145/1007730.1007735.
[17] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *Intern J Pattern Recognit Artif Intell*, vol. 23, no. 4, 2009, doi: 10.1142/S0218001409007326.
[18] G. M. Weiss, "Mining with Rarity: A Unifying Framework," *SIGKDD Explorations*, vol. 6, no. 1, 2004.
[19] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, 2018, doi: 10.1016/j.neunet.2018.07.011.
[20] H. He and Y. Ma, *Imbalanced learning: Foundations, algorithms, and applications*. 2013. doi: 10.1002/9781118646106.