TESIS

PREDIKSI *LUMPY SKIN DISEASE* BERDASARKAN FITUR METEREOLOGI DAN GEOSPASIAL MENGGUNAKAN ALGORITMA *RANDOM FOREST* DAN *SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE*

Disusun oleh:

Nama : **Suparyati**
NIM : **21.55.1023**
Konsentrasi : **Business Intelligence**

PROGRAM STUDI S2 TEKNIK INFORMATIKA

PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2023

TESIS

PREDIKSI *LUMPY SKIN DISEASE* BERDASARKAN FITUR METEREOLOGI DAN GEOSPASIAL MENGGUNAKAN ALGORITMA *RANDOM FOREST* DAN *SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE*

LUMPY SKIN DISEASE PREDICTION BASED ON METEREOLOGICAL AND GEOSPATIAL FEATURES USING RANDOM FOREST ALGORITHM AND SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE

Diajukan untuk memenuhi salah satu syarat memperoleh derajat Magister

Disusun oleh:

Nama : **Suparyati**
NIM : **21.55.1023**
Konsentrasi : **Business Intelligence**

PROGRAM STUDI S2 TEKNIK INFORMATIKA

PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2023

HALAMAN PENGESAHAN

PREDIKSI LUMPY SKIN DISEASE BERDASARKAN FITUR METEREOLOGI DAN GEOSPASIAL MENGGUNAKAN ALGORITMA RANDOM FOREST DAN SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE

LUMPY SKIN DISEASE PREDICTION BASED ON METEREOLOGICAL AND GEOSPATIAL FEATURES USING RANDOM FOREST ALGORITHM AND SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE

Dipersiapkan dan Disusun oleh

Suparyati

21.55.1023

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Teknik Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Selasa, tanggal 04 April 2023

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 04 April 2023
Rektor

Prof. Dr. M. Suyanto, M.M.
NIK. 190302001

HALAMAN PERSETUJUAN

PREDIKSI LUMPY SKIN DISEASE BERDASARKAN FITUR METEREOLOGI DAN GEOSPASIAL MENGGUNAKAN ALGORITMA RANDOM FOREST DAN SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE

LUMPY SKIN DISEASE PREDICTION BASED ON METEREOLOGICAL AND GEOSPATIAL FEATURES USING RANDOM FOREST ALGORITHM AND SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE

Dipersiapkan dan Disusun oleh

Suparyati

21.55.1023

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Teknik Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Selasa, tanggal 04 April 2023

Pembimbing Utama

Prof. Dr. Ema Utami, S.Si., M.Kom.
NIK. 190302037

Anggota Tim Penguji

Dhani Ariatmanto, M.Kom., Ph.D.
NIK. 190302197

Hanafi, S.Kom., M.Eng., Ph.D.
NIK. 190302024

Pembimbing Pendamping

Alva Hendi M., S.T., M.Eng., Ph.D.
NIK. 190302493

Prof. Dr. Ema Utami, S.Si., M.Kom.
NIK. 190302037

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 04 April 2023
Direktur Program Pascasarjana

Prof. Dr. Kusrini, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN TESIS

Yang bertandatangan di bawah ini,

Nama mahasiswa : Suparyati
NIM : 21.55.1023
Konsentrasi : Business Intelligence

Menyatakan bahwa Tesis dengan judul berikut:
Prediksi Lumpy Skin Disease Berdasarkan Fitur Meteorologi dan Geospasial Menggunakan Algoritma Random Forest dan Synthetic Minority Oversampling Technique.

Dosen Pembimbing Utama : Prof. Dr. Ema Utami, S. Si., M. Kom
Dosen Pembimbing Pendamping : Alva Hendi Muhammad, S.T., M.Eng., Ph.D.

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Tim Dosen Pembimbing
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi

Yogyakarta, 04 April 2023
Yang Menyatakan,

Suparyati

HALAMAN PERSEMBAHAN

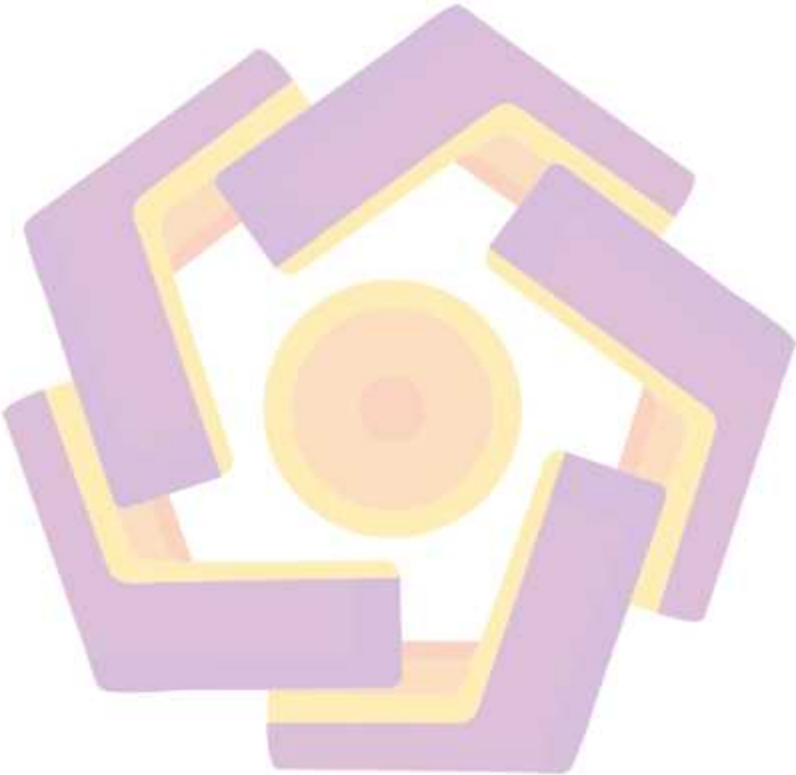
Puji Syukur penulis panjatkan kepada Allah SWT, yang telah memberikan kesehatan, rahmat dan hidayah, sehingga penulis masih diberikan kesempatan untuk menyelesaikan tesis ini, sebagai salah satu syarat untuk mendapatkan gelar Magister. Walaupun jauh dari kata sempurna, namun penulis bangga telah mencapai pada titik ini, yang akhirnya skripsi ini bisa selesai diwaktu yang tepat.

Tesis ini saya persembahkan kepada :

1. Ibu dan Almarhum Bapak, terima kasih atas doa, pengorbanan, nasehat serta kasih sayang yang tidak pernah henti.
2. Suamiku, terimakasih atas kesabaran dan doa serta motivasi yang telah diberikan selama ini.
3. Adik-adik ku terimakasih doa dan semangatnya.
4. Semua teman-teman PJJ Teknik Informatika 2021.
5. Bapak Ibu Pimpinan dan seluruh pegawai pada Balai Besar Veteriner Maros dan Balai Penerapan Standarisasi Instrumen Pertanian Jawa Tengah atas support-nya selama saya menempuh studi.
6. Kepada semua teman-teman, saudara yang tidak bisa saya sebutkan satu persatu, saya persembahkan tesis ini untuk kalian semua.

HALAMAN MOTTO

"Man jadda wa jadda"



KATA PENGANTAR

Puji syukur penulis panjatkan kehadirat Allah SWT yang telah memberikan dan menganugrahkan kasih sayang, rezeki, dan kesehatan serta atas berkah, ridho dan hidayahNya, sehingga saya sebagai penulis dapat menyelesaikan skripsi dengan judul "Prediksi *Lumpy Skin Disease* Berdasarkan Fitur Metereologi dan Geospasial Menggunakan Algoritma *Random Forest* dan *Synthetic Minority Oversampling Technique*". Tesis ini disusun sebagai salah satu syarat untuk memperoleh gelar Magister sekaligus pertanggungjawaban akhir penulis sebagai mahasiswa jurusan PJJ-Magister Teknik Informatika, Universitas Amikom Yogyakarta. Penulis menyadari bahwa dalam penyusunan tesis ini masih ada kekurangan dan kesalahan, maka dari itu, penulis dengan penuh kerendahan hati mengharapkan dan menerima saran dan kritikan dari berbagai pihak untuk dijadikan bahan masukan dan evaluasi dalam perbaikan dan kesempurnaan penulisan tesis ini.

Tesis ini dapat terselesaikan karena adanya kerja keras, tanggung jawab untuk menyelesaikan tesis ini dan tidak terlepas dari doa, bimbingan dan dukungan dari berbagai pihak, serta kritik dan saran yang membantu terselesaikannya penulisan tesis ini. Oleh karena itu, pada kesempatan ini penulis ingin mengucapkan rasa terima kasih yang sebesar-besarnya kepada :

1. Prof. DR. M. Suyanto, MM selaku rektor Universitas Amikom Yogyakarta.
2. Prof. Dr. Kusrini, M.Kom. selaku Direktur Program Pascasarjana dan Ketua Program Studi S2 PJJ Teknik Informatika Universitas Amikom Yogyakarta.

3. Prof. Dr. Ema Utami, S.Si., M.Kom. selaku Wakil Direktur Program Pascasarjana Universitas Amikom Yogyakarta dan sebagai dosen pembimbing utama, terima kasih banyak atas curahan waktu,tenaga dan fikiran serta kesabaran yang telah diberikan dalam membimbing kami sehingga tesis ini dapat terselesaikan dengan baik.
4. Alva Hendi Muhammad, M.Eng., Ph.D. selaku Sekretaris Program Studi S2 PJJ Teknik Informatika Universitas Amikom Yogyakarta dan sebagai dosen pembimbing pendamping, terima kasih banyak atas kesabaran dan bimbingan yang telah diberikan dan kebijaksanaannya berkenan dalam membimbing penulis sehingga tesis ini dapat diselesaikan dengan baik.
5. Dhani Ariatmanto, M.Kom., Ph.D. selaku dosen penguji, terima kasih banyak atas bimbingan, kritik dan sarannya untuk perbaikan tesis ini.
6. Hanafi, S.Kom., M.Eng., Ph.D. selaku dosen penguji yang telah banyak memberikan masukan dan bimbingan dalam skripsi ini.
7. Seluruh Bapak dan Ibu dosen dan Staf Pengelola Program Studi S2 PJJ Teknik Informatika Universitas Amikom Yogyakarta yang telah memberikan bekal ilmu.
8. Bapak Ibu Staf Pengelola Program Studi S2 PJJ Teknik Informatika Universitas Amikom Yogyakarta yang telah banyak membantu kelancaran dalam proses studi.

Bagi seluruh pihak yang tidak bisa penulis sebutkan namanya satu persatu, penulis mengucapkan rasa terima kasih banyak atas segala doa dan dukungannya serta mohon maaf yang sebesar-besarnya. Semoga segala kebaikan, bantuan dan

amal baik dari berbagai pihak tersebut diatas mendapat balasan yang setimpal dari Allah SWT dan penulis senantiasa berharap semoga skripsi yang dibuat ini dapat bermanfaat untuk berbagai pihak.

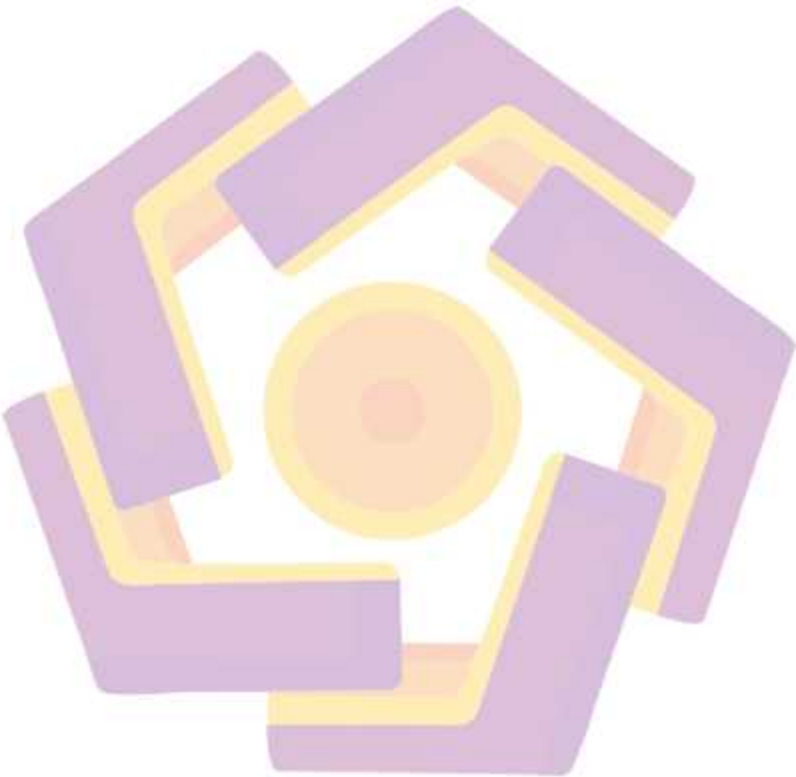
Yogyakarta, 04 April 2023

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR



INTISARI

Lumpy skin disease (LSD) merupakan salah satu penyakit pada sapi yang baru masuk ke Indonesia. Pencegahan dini penyebaran penyakit sangat diperlukan. *Machine Learning* membantu mengklasifikasikan LSD dengan memanfaatkan kumpulan data LSD yang ada dari Mendeley Data. Salah satu permasalahan dalam klasifikasi menggunakan machine learning adalah ketidakseimbangan data sehingga diperlukan teknik *resampling*. Tujuan dari penelitian ini adalah untuk mengoptimalkan klasifikasi Random Forest dalam memprediksi LSD dengan menggunakan *Synthetic Minority Oversampling Technique* (SMOTe) untuk menangani kelas data yang tidak seimbang.

Hasil percobaan yang dilakukan menunjukkan bahwa penggunaan SMOTe pada klasifikasi Random Forest memberikan peningkatan terhadap nilai kinerja model dibandingkan dengan penelitian terdahulu oleh Savafi (2022) yang berupa nilai *Recall* meningkat 27% dari 71% menjadi 98%. F1-*Score* memiliki peningkatan 17% dari 79% menjadi 96% serta AUC meningkat sebesar 13% dari 85% menjadi 98%. Peningkatan nilai akurasi sebesar 3% dari sebelumnya 96% menjadi 99% menunjukkan bahwa model dapat mengklasifikasikan sapi yang tidak terinfeksi dengan baik. Metrik *recall* menjadi perhatian dalam penelitian ini pada klasifikasi LSD dengan harapan semakin tinggi skor, kesalahan klasifikasi sapi yang terinfeksi diprediksi oleh model menjadi sehat.

Masih ada peluang untuk meningkatkan *recall* dengan menggunakan berbagai teknik *resampling* dan menentukan parameter model yang tepat di awal.

Kata kunci: *Imbalanced Data, Lumpy Skin Disease, Machine Learning, Random Forest, SMOTe.*

ABSTRACT

Lumpy skin disease (LSD) is a disease in cattle that has just entered Indonesia. Early prevention of the spread of disease is essential. Machine Learning helps classify LSD by leveraging existing LSD datasets from Mendeley Data. One of the problems in classification using machine learning is to classify data so that a resampling technique is needed. The purpose of this research is to optimize the Random Forest classification in predicting LSD by using the Synthetic Minority Oversampling Technique (SMOte) to handle class unbalanced data.

The results of the experiments conducted showed that the use of SMOTe in the Random Forest classification provided an increase in the performance value of the model compared to previous research by Savafi (2022) in which the Recall value increased 27% from 71% to 98%.. F1-Score increased by 17% from 79% to 96% and AUC increased by 13% from 85% to 98%. An increase in the accuracy value of 3% from the previous 96% to 99% indicates that the model can classify cattle that are not infected properly. Recall metrics are of concern in this study on LSD classification with the hope that the higher the score, the misclassification of cattle predicted by the model will be healthy.

There are still opportunities to improve recall by using various resampling techniques and setting the right model parameters up front.

Keyword: Imbalanced Data, Lumpy Skin Disease, Machine Learning, Random Forest, SMOTe.

BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

Lumpy Skin Disease (LSD) merupakan penyakit pada sapi, yang disebabkan oleh virus dari genus *Capripox*, famili *Poxviridae* (Lojkic, 2018). Penyakit ini dapat menginfeksi sapi dan kerbau yang ditandai dengan adanya nodul-nodul di tubuh sapi, demam, nafsu makan menurun sehingga menyebabkan tubuh ternak kurus (Garcia, 2017). LSD ditemukan pertama kali pada tahun 1929 di Afrika (Morris et al, 1930). Penyakit ini telah menyebar melalui Afrika, Timur Tengah, Eropa Tenggara, Asia dan Cina yang sekarang menjadi endemik di banyak negara Afrika, serta wilayah Timur Tengah (Irak, Arab Saudi, dan Republik Arab Suriah) dan Turki (Dimitriadis, 2020).

Penyakit ini telah mengakibatkan kerugian ekonomi yang besar di negara-negara yang terdampak seperti penurunan produksi susu dan daging sapi yang cukup besar. Konsekuensi lain dari penyakit ini termasuk kerusakan kulit, penurunan laju pertumbuhan sapi potong, infertilitas sementara atau seumur hidup, aborsi, kematian pada hewan yang terinfeksi, biaya pengobatan simptomatis ternak yang terinfeksi dan vaksinasi, biaya pencegahan kontak vektor dan hewan, disinfeksi lokasi ternak, serta biaya kompensasi apabila menerapkan sistem *stamping out* (Alemayehu et al., 2013), (Namazi & Tafti, 2021),(Sendow et al., 2021).

Faktor risiko terjadinya infeksi LSD diantaranya adalah kondisi lingkungan, letak demografi, manajemen peternakan, populasi vektor, dan data epidemiologi termasuk pergerakan hewan, virulensi virus, status imun, iklim baik angin dan curah hujan (Pande, 2019) (Ince, 2020). Daerah basah dan lembab merupakan resiko tinggi bagi populasi sapi karena merupakan tempat berkembangbiaknya vector mekanik pembawa virus LSD seperti Aedes sp. Peluang penularan virus LSD akan meningkat seiring dengan peningkatan vector mekanik yang menyebabkan prevalensi LSD akan mengalami peningkatan (Molla et al., 2018). Prevalensi LSD menjadi lebih tinggi pada daerah dengan curah hujan tahunan rata-rata> 1000 mm (Ochwo et al., 2019). Kondisi iklim memiliki hubungan langsung dengan kelangsungan hidup vektor yang memainkan peran penting dalam epidemiologi penyakit. Iklim yang hangat dan lembab, kondisi lingkungan yang mendukung masuknya populasi vektor, seperti yang terlihat selama hujan musiman, dan pengenalan hewan baru ke kawanan, semuanya merupakan faktor risiko penyebaran virus LSD. Selain itu, arah dan intensitas angin berperan dalam penyebaran virus (Chihota, 2003).

Hubungan antara infeksi LSD dan faktor meteorologi dan geospasial telah dipelajari dalam banyak penelitian yang menemukan bahwa faktor-faktor seperti suhu, curah hujan, luasan lahan, kelembaban, dan kecepatan angin dapat memprediksi atau mempengaruhi terjadinya penyakit (Machado et al., 2019) , (Alkhamis, 2016), (Allepuz et al., 2019); (Machado et al., 2019); (Molla et al., 2017) (Sprygin et al., 2019), (Sprygin et al., 2019), (Tuppurainen & Oura, 2012).

Indonesia yang memiliki iklim tropis dengan curah hujan tinggi serta letak demografi yang ada memungkinkan terjadinya penyebaran virus LSD.

Prediksi dini terhadap keberadaan LSD sangatlah penting dalam upaya pencegahan terjadinya penyebaran virus tersebut. Dari tahun ke tahun adanya virus LSD semakin meluas ke berbagai negara, termasuk Indonesia sendiri dimana virus Kementerian Pertanian Republik Indonesia melalui Direktorat Jenderat Peternakan dan Kesehatan Hewan telah melaksanakan berbagai upaya pencegahan masuknya penyakit LSD ini ke Indonesia sejak penyakit ini masuk ke Asia Tenggara sejak tahun 2019. LSD pertama kali menginfeksi ternak sapi di Provinsi Riau pada bulan Februari 2022. Langkah pengamanan pencegahan penyebaran LSD dilakukan dengan cara vaksinasi, yang harus didukung dengan deteksi dini dan penelusuran kasus, pengendalian lalu lintas, pengendalian vektor, serta komunikasi, informasi dan edukasi (Pertanian, 2022).

Seiring dengan perkembangan teknologi informasi, produksi data yang dapat dilakukan oleh siapa saja dan dimana saja mengakibatkan banyaknya data dalam jumlah yang sangat melimpah. Kebutuhan suatu metode untuk merepresentasikan data-data tersebut ke dalam suatu informasi berguna menjadi hal yang sangatlah penting. Dalam bidang ilmu komputer, *machine learning* sudah sejak lama digunakan untuk membantu proses pengklasifikasian terhadap berbagai permasalahan yang ada. Hal ini juga dapat dimanfaatkan dalam bidang kesehatan hewan seperti pendeteksian penyakit LSD pada sapi (Safavi, 2022),(Rai et al., 2020), prediksi penyakit mastitis pada ternak (Ghafoor, 2021)(Mammadova & Keskin, 2013)(Hyde et al., 2020), penyakit kulit pada ternak (Workee, 2021),

resiko penyakit pernafasan pada sapi (Rojas et al., 2022), penyakit *postpartum* pada sapi perah (Beek et al., 2018) serta deteksi otomatis kepincangan pada domba (Kaler et al., 2020). Dengan banyaknya penelitian yang telah memanfaatkan teknik *machine learning* dalam melakukan prediksi terhadap suatu penyakit, maka menjadi dasar bagi penulis untuk dapat melakukan penelitian yang sama dengan melakukan berbagai modifikasi untuk dapat menghasilkan prediksi dengan akurasi yang optimal dan hasil yang tidak bias.

Algoritma *machine learning* dapat menghasilkan klasifikasi bias ketika menghadapi *dataset* yang tidak seimbang. Kondisi *imbalanced data* dapat terlihat secara nyata pada himpunan data yang memiliki dua kelas. Kelas yang jumlah *instance* terkecil (*minority class)* dan kelas yang jumlah *instance* terbesar (*majority class)*. Algoritma klasifikasi biasa cenderung hanya akan memprediksi kelas data mayoritas. Fitur pada kelas data minoritas akan dianggap *noise* sehingga akan diabaikan dalam proses klasifikasinya sehingga memungkinkan terjadinya kesalahan klasifikasi dari kelas minoritas. Penanganan *imbalance dataset* dilakukan dengan teknik resampling. Dalam penelitian prediksi resiko penyakit dengan dataset yang sangat tidak seimbang, untuk Teknik *resampling*-nya menggunakan metode *Repeat Random Sub-sampling* (Khalilia et al., 2011). Teknik *Synthetic Minority Oversampling Technique* (SMOTe) telah digunakan pada dataset yang tidak seimbang dalam penelitian mengenai prediksi penyebaran COVID-19 (Aljameel et al., 2021). Selain itu teknik *random under sampling, SMOeE* dan *SMOTe-Tomek* dalam penanganan *imbalanced dataset* telah dilakukan pada dataset yang diperoleh dari twitter (Utami et al., 2021). Penanganan dataset yang tidak

seimbang sangatlah diperlukan dalam penelitian yang akan dilakukan, sehingga peneliti akan menggunakan *Synthetic Minority Oversampling Technique* (SMOTe) untuk menyeimbangkan *dataset* LSD.

1.2. Rumusan Masalah

Rumusan masalah dalam penelitian ini yang didapatkan dari latar belakang masalah tersebut diatas sebagai berikut:

a. Bagaimana mengimplementasikan *SMOTe* pada dataset yang digunakan dalam proses klasifikasi Random Forest untuk memprediksi keberadaan LSD?
b. Apakah *Synthetic Minority Oversampling Technique* (SMOTe) yang diterapkan pada dataset LSD dapat mengatasi imbalanced dataset?
c. Apakah metode *oversampling SMOTe* yang digunakan dapat mengoptimasi kinerja *Random Forest Classifier*?
d. Apakah model yang dirancang dapat memberikan kinerja yang baik (*finiteness*) dalam merespon dataset yang dievaluasi?
e. Berapa nilai performa (akurasi, *Recall*, Presisi, *f1 score*) yang dihasilkan setelah dilakukan pengujian terhadap model yang telah dilatih menggunakan *dataset lumpy skin disease*?

1.3. Batasan Masalah

Batasan masalah dalam penelitian ini adalah sebagai berikut:

a. Data yang digunakan adalah dataset *lumpy skin disease (LSD)* yang diambil dari Mendeley Data;
b. Teknik *resampling* terhadap dataset *LSD* menggunakan *Synthetic Minority Oversampling Technique (SMOTe)*;

c. Pembagian data menjadi *train set* dan *test set* dengan perbandingan 20:80;

d. Pemodelan data dilakukan menggunakan Bahasa pemrograman *Phyton* dengan *platform Google Colabs*;

e. Pengukuran kinerja model menggunakan akurasi, *recall*, presisi, dan *f1 score*.

1.4. Tujuan Penelitian

Tujuan yang ingin dicapai dari penelitian ini sebagai berikut:

a. Mengatasi ketidakseimbangan kelas data pada *dataset* LSD;

b. Mengimplementasikan teknik *resampling* SMOTe terhadap dataset LSD;

c. Mengoptimasi *Random Forest Classifier* dalam memprediksi keberadaan LSD;

d. Mengetahui nilai kinerja model berdasarkan evaluasi melalui akurasi, presisi, *recall*, dan *f1 score*.

1.5. Manfaat Penelitian

Manfaat yang didapatkan dalam penelitian yang dilakukan sebagai berikut:

a. Dapat menjadi pedoman pengembangan penelitian dalam mendeteksi keberadaan virus LSD dengan *optimasi* pada algoritma Random Forest serta penerapan SMOTe terhadap dataset LSD;

b. Berkontribusi secara ilmiah terhadap pengembangan penelitian di bidang *Machine Learning*, khususnya pendeteksian berdasarkan fitur meteorologi dan geospasial;

c. Hasil kinerja model yang diajukan diharapkan dapat menjadi bahan pertimbangan untuk dapat dikembangkan lebih lanjut dalam bidang kecerdasan buatan dalam pendeteksian keberadaan LSD pada ternak.

BAB II

TINJAUAN PUSTAKA

2.1. Tinjauan Pustaka

Sejauh pengetahuan peneliti, dalam bidang machine learning untuk melakukan prediksi terhadap *Lumpy Skin Disease* pada ternak baru terdapat dua buah penelitian Penelitian pertama mengenai pendeteksian LSD yang dilakukan menggunakan metodologi *Deep Convolutional Neural Network* memiliki akurasi 92,5% untuk mendeteksi *Lumpy skin* dan *normal skin* (Rai et al., 2020). Sedangkan penelitian kedua mengenai prediksi keberadaan LSD berdasarkan fitur meteorologi dan geospatial telah dilakukan dengan memanfaatkan *machine learning* yang menghasilkan akurasi 97% pada penggunaan model *Artificial Neural Network* (Safavi, 2022). Kedua penelitian tersebut di atas menggunakan dua jenis *dataset* yang berbeda. Pada penelitian pertama menggunakan jenis gambar yang terinfeksi LSD dan normal, sedangkan penelitian kedua menggunakan jenis file .csv. Adapun penelitian yang akan dilakukan mengadopsi dataset yang telah digunakan pada penelitan kedua, dengan melakukan teknik *resampling* terlebih dahulu untuk datasetnya.

Penggunaan *machine learning* dalam melakukan prediksi risiko mastitis pada sapi menggunakan algoritma random forest, dapat memprediksi risiko mastitis pada sapi dengan akurasi 98,8% (Ghafoor, 2021). Sedang dalam penelitian lainnya mengenai prediksi otomatis pada mastitis sapi, *Random Forest* memberikan hasil perfoma yang sangat baik dibandingkan dengan algoritma yang lainnya (Hyde et

al., 2020). *Machine learning* juga berperan dalam melakukan prediksi terhadap penyakit kulit sapi dengan menggunakan perbandingan tiga teknik filter (*median, gaussian*, dan *gabor filter* (Workee, 2021). Prediksi pada resiko penyakit pernafasan penggemukan sapi, pada area *Under the Curve, Random Forest Model* memiliki performa terbaik sebesar 0.789 menggunakan *testing dataset* (Rojas et al., 2022). Dalam melakukan prediksi penyakit postpartum pada sapi perah, *Random Forest* juga telah dibuktikan sebagai metode yang terbaik (Beek et al., 2018). Sedangkan pada wabah *African swine fever* di seluruh dunia menggunakan variable bioklimat didapatkan hasil bahwa algoritma R*andom Forest* mengungguli teknik lain dengan akurasi 80,4% dalam dataset yang berisi semua variabel prediktif, dan algoritma SVM menunjukkan akurasi terbaik dalam *subset dataset* yang hanya berisi fitur iklim penting (76,02%) (Liang et al., 2020). Berdasarkan penelitian-penelitian yang telah dilakukan di atas memberikan gambaran bahwa algoritma *Random Forest* memiliki hasil yang lebih baik dibandingkan dengan penggunaan algoritma yang lainnya dalam melakukan prediksi terhadap penyakit pada hewan.

Sebelum dilakukan klasifikasi, resampling terhadap dataset perlu dilakukan untuk menangani data yang tidak seimbang. Teknik *resampling* menggunakan *synthetic minority oversampling technique (SMOTE)* telah digunakan dalam menyeimbangkan dataset yang digunakan untuk melakukan prediksi tingkat keparahan penyakit Covid-19 (Aljameel et al., 2021), analisa temperamen Keirsey (Iskandar et al., 2020), prediksi tingkat keselamatan pasien gagal jantung (Rahayu et al., 2020). SMOTE merupakan teknik *oversampling* populer yang menghasilkan

kumpulan data sintetis baru di sekitar sampel minoritas. Data sintetik untuk kelas minoritas dihasilkan dengan melakukan interpolasi disekitar tetangga terdekat dari kelas minoritas yang berurutan (Bellinger et al., 2017).

Merujuk pada penelitan-penelitian yang telah dilakukan, maka dalam penelitian ini akan menggunakan teknik resampling SMOTE yang diterapkan pada *Random Forest Classifier* untuk memprediksi penyakit *Lumpy Skin Disease*.

2.2. Keaslian Penelitian

Tabel 2.1. Matriks literatur review dan posisi penelitian

Prediksi *Lumpy Skin Disease* Berdasarkan Fitur Metereologi dan Geospasial Menggunakan Algoritma Random Forest dan *Synthetic Minority Oversampling Technique*

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
1	*Assessing machine learning techniques in forecasting lumpy skin disease occurrence based on meteorological and geospatial features* (Safavi, 2022)	Ehsanallah Afshari Safavi. Tropical Animal Health and Production (2022) 54:55	Membangun model prediksi LSD menggunakan beberapa algoritma ML berdasarkan meteorologi dan gespasial.	ANN dengan akurasi 97% dapat digunakan untuk memprediksi keberadaan virus LSD dengan parameter meteorologi dan gespasial.	Tidak dijelaskan *preprocessing* terhadap *imbalance dataset* atau jumlah kasus terinfeksi LSD serta banyak *missing value* pada dataset yang digunakan.	Dilakukan resampling dengan SMOTE pada dataset LSD. Algoritma menggunakan Random Forest
2	*A Deep Learning Approach to Detect Lumpy Skin Disease in Cows* (Rai et al., 2020)	Gaurav Rai, Naveen, Aquib Hussain, Amit Kumar and Rahul Nijhawan. EasyChair 2020	Deteksi LSD berdasarkan gambar *lumpy skin* dan *normal skin* menggunakan *Deep Convolutional Neural Network*.	Prediksi LSD menggunakan *Deep Convolutional Neural Network* memiliki akurasi 92,5%.	Belum adanya standarisasi dataset berupa gambar yang terinfeksi LSD dengan terinfeksi penyakit kulit yang hampir mirip.	Penelitian sebelumnya menggunakan *Deep Convolutional Neural Network. sedangkan penelitian yang akan dilakukan menggunakan Random Forest Classifier*

Tabel 2.1. (Lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
3	*A Machine Learning Application to Predict Risk of Mastitis in Cattle from AMS Sensor Data* (Ghafoor, 2021)	Naeem Abdul Ghafoor, Beata Sitkowska. Agriengineering 2021	Menentukan parameter paling efektif yang dapat digunakan untuk memprediksi risiko mastitis pada sapi	MasPA, yang didasarkan pada algoritma *random forest*, dapat memprediksi risiko mastitis pada sapi dengan akurasi 98,8%.	Aplikasi MasPA berbasis web dan standalone. Ke depannya dapat dibuat *open source* sehingga dapat diintegrasikan ke dalam AMS untuk mendeteksi risiko mastitis secara real time.	Hyperparameter tunning yang digunakan *GridSearch* untuk optimalisasi. Sedangkan dalam penelitian yang akan dilakukan akan menggunakan *TPOT* untuk optimasinya.
4	*Automated detection of lameness in sheep using machine learning approaches: novel insights into behavioural differences among lame and non-lame sheep* (Kaler et al., 2020)	Jasmeet Kaler et. al Royal Society Open Science 2020	Mengembangkan dan membandingkan algoritma yang dapat membedakan ketimpangan dalam tiga aktivitas berbeda (berjalan, berdiri, dan berbaring).	Algoritma *random forest* bekerja paling baik untuk mengklasifikasi kepincangan dengan akurasi 84,91% dalam berbaring, 81,15% dalam berdiri dan 76,83% dalam berjalan dan secara keseluruhan diklasifikasikan dengan benar lebih dari 80% dalam aktivitas domba	Perbedaan perilaku baru antara domba lumpuh dan tidak lumpuh di ketiga aktivitas dapat digunakan untuk mengembangkan sistem otomatis untuk deteksi ketimpangan.	Penelitian yang telah dilakukan menggunakan random forest yang optimasi algoritmanya tidak disebutkan. Penelitian yang akan dilakukan menggunanakan *TPOT* untuk optimasi *Random Forest*.

Tabel 2.1. (Lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
5.	*Automated prediction of mastitis infection patterns in dairy herds using machine learning* (Hyde et al., 2020)	R. Hyde et. Al Scientific Reports 2020	Melakukan Prediksi otomatis terhadap pola infeksi mastitis pada peternakan sapi perah menggunakan pembelajaran mesin	Random Forest memberikan kinerja model terbaik yang dinilai berdasarkan akurasi, PPV, dan NPV	Subsampling yang tidak digunakan untuk memperkirakan model akhir dikarenakan tidak meningkatkan kinerja model sangat disayangkan, karena akurasi yang didapat bisa saja bias apabila datasetnya tidak seimbang.	Dalam memperkirakan model akhir tidak menggunakan subsampling serta tidak dijelaskan apakah ada optimasi yang dilakukan pada random forest. Penelitian yang akan datang akan melakukan SMOTE terhadap dataset serta menggunakan Random Forest.
6	*Cattle skin diseases identification model using machine learning approach* (Workee, 2021)	Workee, Getachew Bahir Dar University 2021	Melakukan identifikasi penyakit kulit sapi menggunakan perbandingan tiga teknik filter (median, gaussian, dan gabor filter)	Akurasi yang dicapai 96,5% di CNN, 93% dengan HOG, dan 98,75% menggunakan fitur hybrid.	Kedepannya dapat menerapkan deep learning untuk teknik filtering yang tepat guna menghilangkan bulu hewan untuk memecah bagian terpenting dari bagian yang sakit menuju peningkatan kemampuan diskriminasi fitur algoritma pada penyakit kulit hewan	Penelitian yang telah dilakukan dalam bidang *computer vision* untuk pengenalan penyakit kulit, dan klasifikasi. Penelitian yang akan dilakukan menggunakan data tabular bukan *image*.

Tabel 2.1. (Lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
7	*Predicting Bovine Respiratory Disease Risk in Feedlot Cattle in the First 45 Days Post Arrival* (Rojas et al., 2022)	Hector A. Rojas, et. al MDPI 2022	Memprediksi Risiko Penyakit Pernafasan Sapi pada penggemukan Sapi	Model prediktif dapat berguna untuk menggambarkan ternak sebagai sapi yang berisiko tinggi atau rendah terhadap penyakit dan dapat memberikan nilai ekonomi relatif terhadap metode standar.	Diperlukannya variabel prediktor dan pengumatan data untuk terus menyempurnakan algoritma dan memberikan perkiraan yang lebih baik dari kinerja prediktif setiap model.	Penelitian yang telah dilakukkan berfokus pada penyakit pernafasan sapi yang menggunakan algoritma *Naïve Bayes, Decision Tree, Random Forest serta Logistic Regression.* Penelitian yang akan dilakukan berfokus pada LSD dengan menggunakan *Random Forest*
8	*Prediction of postpartum diseases of dairy cattle uxing machine learning* (Beek et al., 2018)	S. Beek, et. al Proceedings of the World Congress on Genetics Applied to Livestock Production 2018	Memprediksi penyakit postpartum pada sapi perah menggunakan *machine learning*	Probabilitas penyakit *postpartum* diprediksi berdasarkan informasi *prepartum* yang diberikan untuk algoritma *Random Forest* dengan kepastian yang relatif baik.		Pada penelitian ini dilakukan prediksi terhadap penyakit postpartum pada ternak. Penelitian yang akan dilakukan akan melakukan prediksi terhadap keberadaan LSD suatu wilayan menggunakan algoritma yang sama.
9	*Prediction for global African swine fever outbreaks based on a combination of*	Ruirui Liang et. al. Transboundary and emerging diseases 2020	Memrediksi penyakit ASF dengan Random Forest dan	Algoritma Random Forest mengungguli teknik lain dengan akurasi 80,4% dalam	Masih dapat dilakukan optimasi algoritma untuk meningkatkan akurasi *machine*	Pada penelitian terdahulu, algoritm *Random Forest* dilakukan setelah metode *feature selection*, sedangkan

Tabel 2.1. (Lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
	random forest algorithms and meteorological Data (Liang et al., 2020)		kombinasi data meteorologi	dataset yang berisi semua variabel prediktif, dan Algoritma SVM menunjukkan akurasi terbaik dalam subset dataset yang hanya berisi fitur iklim penting (76,02%).	*learning* yang di ajukan.	pada penelitian yang akan dilakukan penggunaan *Random Forest* dan teknik SMOTe untuk penanganan *imbalanced data*-nya

2.3. Landasan Teori

Adapun dalam penelitian ini, terdapat beberapa landasan teori terkait prediksi yang akan dilakukan yaitu *Lumpy Skin Disease*, *Machine Learning*, *Random Forest Classifier*, *Feature Selection*, dan *Synthetic Minority Oversampling Technique*.

2.3.1. *Lumpy Skin Disease*

Lumpy Skin Disease (LSD) merupakan penyakit pada sapi, yang disebabkan oleh virus dari genus *Capripox*, famili *Poxviridae* (Lojkic, 2018). Penyakit ini dapat menginfeksi sapi dan kerbau yang ditandai dengan adanya nodul-nodul di tubuh sapi, demam, nafsu makan menurun sehingga menyebabkan tubuh ternak kurus (Garcia, 2017). Kadang-kadang hewan yang terinfeksi hanya mengembangkan beberapa benjolan, tetapi dalam kasus yang parah, nodul kulit dapat menutupi seluruh tubuh. Sapi yang terkena menunjukkan cairan mata dan hidung, dan lesi ulseratif dapat dideteksi pada selaput lendir mulut, hidung dan mata (Epizooties, 2017). Sapi yang terinfeksi menunjukkan pembesaran kelenjar getah bening *subscapular* dan *precrural* (Tuppurainen et al., 2018). LSD ditemukan pertama kali pada tahun 1929 di Afrika (Morris et al. 1930). Penyakit ini telah menyebar melalui Afrika, Timur Tengah, Eropa Tenggara, Asia dan Cina yang sekarang menjadi endemik di banyak negara Afrika, serta wilayah Timur Tengah (Irak, Arab Saudi, dan Republik Arab Suriah) dan Turki (Dimitriadis, 2020).

2.3.2. *Machine Learning*

Bidang *machine learning* telah menerima beberapa definisi formal dalam literatur. Arthur Samuel mendefinisikan *machine learning* sebagai bidang studi

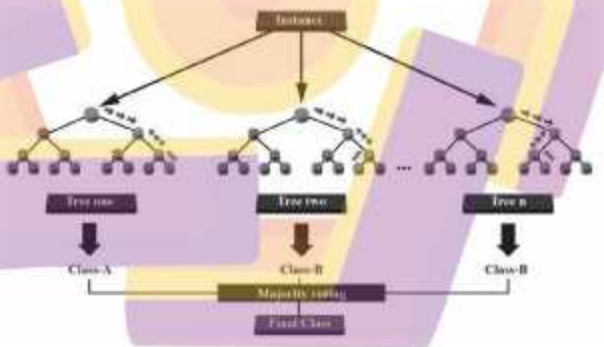
yang memberi komputer kemampuan untuk belajar tanpa diprogram secara eksplisit (Samuel, 2000). Sedangkan menurut Tom Mitchell, *machine learning* merupakan sebuah program komputer yang belajar dari pengalaman yang berhubungan dengan beberapa tugas dan *performance* (Jordan & Mitchell, 2015). *Machine learning* yang merupakan teknologi pengembangan algoritma komputer yang mampu meniru kecerdasan manusia (Naqa & Murphy, 2015). Dari berbagai definisi tersebut dapat diartikan melatih komputer untuk secara cerdas melakukan tugas di luar penghitungan angka tradisional dengan mempelajari lingkungan sekitar melalui contoh yang berulang.

Machine learning dapat dibagi menurut sifat pelabelan data menjadi *supervised, unsupervised*, dan *semi supervised*. *Supervised learning* digunakan untuk memperkirakan pemetaan yang tidak diketahui (*input, output*) dari sampel yang diketahui (*input, output*), di mana output diberi label (misalnya, klasifikasi dan regresi) (Naqa & Murphy, 2015). Dalam *unsupervised learning*, hanya sampel masukan yang diberikan ke sistem *learning* (misalnya, pengelompokan dan estimasi fungsi kepadatan probabilitas). Sedangkan *semi-supervised learning* yang merupakan kombinasi dari keduanya dimana sebagian data diberi label sebagian dan bagian berlabel digunakan untuk menyimpulkan bagian yang tidak berlabel (misalnya, sistem pencarian teks/gambar).

2.3.3. Random Forest

Random Forest (untuk klasifikasi dan regresi) (Carrizosa et al., 2021), (L. E. O. Breiman, 2001), yang merupakan bentuk pengembangan dari *Decision Tree* (L. Breiman et al., 2017), adalah algoritma *machine learning* yang sangat kuat yang

terdiri dari sejumlah besar jumlah *decision tree* yang beroperasi sebagai *ensambel*. *Random Forest* adalah *classifier* di mana proses pelatihan dilakukan dengan menggunakan metode "*bagging*" (Taser, 2021), (Liaw dan Wiener, 2002). *Random Forest* menghasilkan beberapa *decision tree* dan mengintegrasikannya untuk mendapatkan prediksi yang akurat dan stabil. Setiap *decision tree* belajar dari *random sampling* dari kumpulan data. Sampel diambil dengan penggantian, yang disebut *bootstrap*, artinya beberapa sampel akan digunakan beberapa kali dalam *decision tree* individu. Sebuah sampel dikategorikan ke dalam kelas yang dapat memenangkan suara mayoritas dari keseluruhan *decision tree* di dalam *forest* (Gambar 2.1). Estimasi tak bias dari akurasi prediksi dapat diperoleh berdasarkan data pelatihan yang pada setiap iterasi *bootstrap* kira-kira 1/e sampel pelatihan ditinggalkan sebagai data *out-of-bag* (Liaw & Wiener, 2002).



Gambar 2.1. Flowchart klasifikasi menggunakan algoritma Random Forest

Random Forest dimulai dengan pemurnian *node* anak. Langkah ini dilakukan melalui pemisahan variabel target menurut variabel prediktor dari *node* induk. Prosedur ini berlanjut sampai kriteria berhenti yang telah ditentukan

tercapai. Dengan ini diperoleh klasifikasi atau model regresi sederhana untuk setiap *decision tree*. Akhirnya, nilai rata-rata dari hasil *decision tree* yang berbeda dihitung untuk mencapai model *Random Forest* akhir. Dengan cara ini, ada tiga *hyperparameter* yang dapat diatur sebagai berikut (L. E. O. Breiman, 2001):

[1] *Number of trees* (Jumlah Pohon): Parameter ini menentukan jumlah *decision tree* di *forest* model. *Decision tree* tambahan biasanya dapat meningkatkan akurasi model, karena prediksi dilakukan dengan menggunakan sejumlah besar suara dari *decision tree* yang beragam, meskipun, sejumlah besar *decision tree* menyebabkan peningkatan waktu komputasi;

[2] *Number of split (NS): Parameter NS* dapat mengontrol jumlah minimum sampel yang diperlukan untuk membagi simpul daun internal. Nilai yang terlalu besar dapat menyebabkan under-fitting karena *decision tree* tidak akan dapat membagi waktu yang cukup untuk mencapai kemurnian node;

[3] *Depth (d): Depth* dalam *decision tree* mewakili panjang setiap *decision tree* di mana semakin dalam *decision tree*, semakin banyak perpecahan yang dimilikinya. Setiap *decision tree* dalam model *Random Forest* menghasilkan banyak pemisahan untuk mengisolasi kelas hasil yang homogen. Jumlah split yang lebih besar memungkinkan *decision tree* untuk dideskripsikan.

Tahap utama untuk mencapai prediksi yang akurat dilakukan dengan menyesuaikan parameter yang digunakan untuk melatih model (Wu et al., 2007). Dalam mendefinisikan parameter optimal dengan tujuan meningkatkan akurasi suatu model tidak ada aturan pasti. Diberbagai penelitian, beberapa prosedur

percobaan dan kesalahan empiris direkomendasikan untuk menyesuaikan *hyperparameter* secara optimal.

2.3.4. *Features Selection*

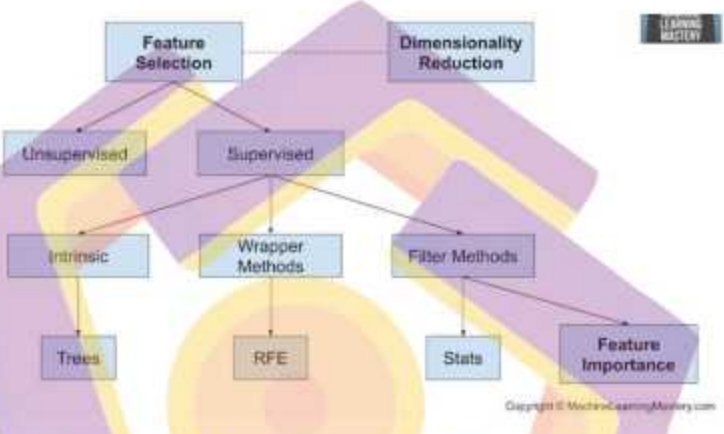
Seleksi Fitur adalah metode untuk mengurangi variabel input ke model dengan hanya menggunakan data yang relevan dan menghilangkan noise dalam data. Beberapa masalah pemodelan prediktif memiliki sejumlah besar variabel yang dapat memperlambat pengembangan dan pelatihan model dan memerlukan sejumlah besar memori sistem. Selain itu, kinerja beberapa model dapat menurun ketika memasukkan variabel input yang tidak relevan dengan variabel target (García et al., 2015). Pemilihan fitur juga terkait dengan dimensi teknik reduksi di mana kedua metode mencari lebih sedikit variabel input ke model prediktif. Perbedaannya adalah bahwa pemilihan fitur memilih fitur untuk disimpan atau dihapus dari kumpulan data, sedangkan pengurangan dimensi membuat proyeksi data yang menghasilkan fitur input yang sama sekali baru. Dengan demikian, pengurangan dimensi adalah alternatif untuk pemilihan fitur daripada jenis pemilihan fitur.

Model pemilihan fitur terdiri dari dua jenis:

1. *Supervised Models*, Pemilihan fitur yang diawasi mengacu pada metode yang menggunakan kelas label keluaran untuk pemilihan fitur dan menggunakan variabel target untuk mengidentifikasi variabel yang dapat meningkatkan efisiensi model

2. *Unsupervised Models*, Pemilihan fitur tanpa pengawasan mengacu pada metode yang tidak memerlukan kelas label keluaran untuk pemilihan fitur diggunakan untuk data yang tidak berlabel.

Gambar 2.2 berikut memberikan ringkasan hierarki teknik pemilihan fitur.



Sumber: MachineLearningMastery.com

Gambar 2.2. Ikhtisar Teknik Seleksi Fitur

Salah satu cara melakukan seleksi fitur dengan matriks korelasi. Matriks korelasi merupakan sebuah tabel yang menampilkan koefisien korelasi untuk variabel yang berbeda. Matriks menggambarkan korelasi antara semua kemungkinan pasangan nilai dalam sebuah tabel. Ini adalah alat yang ampuh untuk meringkas kumpulan data besar dan untuk mengidentifikasi dan memvisualisasikan pola dalam data yang diberikan. Matriks korelasi terdiri dari baris dan kolom yang menunjukkan variabel. Setiap sel dalam tabel berisi koefisien korelasi (Ramsay et al., 1984). Sedangkan *correlation heatmaps* adalah jenis plot yang memvisualisasikan kekuatan hubungan antara variabel numerik. Plot korelasi

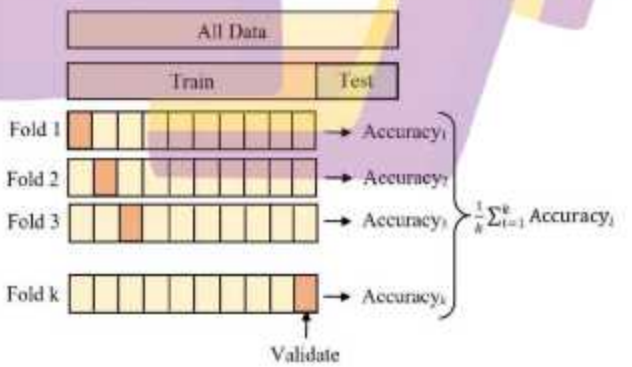
digunakan untuk memahami variabel mana yang terkait satu sama lain dan kekuatan hubungan ini. Plot korelasi biasanya berisi sejumlah variabel numerik, dengan setiap variabel diwakili oleh kolom. Baris mewakili hubungan antara setiap pasangan variabel. Nilai dalam sel menunjukkan kekuatan hubungan, dengan nilai positif menunjukkan hubungan positif dan nilai negatif menunjukkan hubungan negatif yang dapat digunakan untuk menemukan hubungan potensial antara variabel dan untuk memahami kekuatan hubungan ini. Selain itu, plot korelasi dapat digunakan untuk mengidentifikasi outlier dan untuk mendeteksi hubungan linier dan nonlinier. Pengkodean warna sel memudahkan untuk mengidentifikasi hubungan antar variabel secara sekilas. *Correlation heatmaps* dapat digunakan untuk menemukan hubungan linier dan nonlinier antar variabel.

2.3.5. *Cross Validation*

Cross-validation yaitu metode statistik yang digunakan untuk memperkirakan keakuratan model *machine learning*. Setelah model dilatih, belum dapat memastikan seberapa baik model itu akan bekerja pada data yang belum pernah ditemui sebelumnya sehingga diperlukan kepastian mengenai keakuratan kinerja prediksi model. Dalam mengevaluasi kinerja model *machine learning*, beberapa data yang tidak terlihat diperlukan untuk pengujian. Berdasarkan kinerja model pada data yang tidak terlihat, maka dapat ditenntukan apakah model tersebut *underfitting*, *overfitting*, atau digeneralisasi dengan baik. *Cross-validation* dianggap sebagai teknik yang sangat membantu untuk menguji seberapa efektif model *machine learning* ketika data yang ada terbatas. *Cross-validation* akan dilakukan dengan cara sebagian data harus disisihkan untuk pengujian dan validasi;

subset ini tidak akan digunakan untuk melatih model, melainkan disimpan untuk digunakan nanti.

Cross Validation K-fold salah satu metode validasi model dimana data dikategorikan kedalam K *subset* yang sama, yaitu *fold*. Satu *subset* dipertahankan untuk validasi dan *training* dilakukan menggunakan K-1 *fold* lainnya. *Training* terjadi K kali sampai setiap *subset* telah digunakan satu kali untuk validasi *dataset*. Proses ini dapat membuat representasi kesalahan yang lebih baik di seluruh *dataset*, karena semua sampel dikontribusikan baik sebagai pelatihan maupun validasi (Rodriguez et al., 2010). Dalam *K-Fold Cross-validation*, parameter K menunjukkan jumlah lipatan atau bagian yang dibagi menjadi kumpulan data tertentu. Salah satu *fold* dipertahankan sebagai kumpulan validasi dan model *machine learning* dilatih menggunakan K-1 *Fold* yang tersisa. Setiap lipatan *K-Folds* digunakan sebagai set validasi di beberapa titik, dengan skor K (akurasi) sebagai hasilnya. yang kemudian dibuat rata-rata model terhadap setiap lipatan untuk mendapatkan skor akhir untuk model, seperti yang pada Gambar 2.3.



Gambar 2.3. *Cross-validation*

2.3.6. *Imbalanced Dataset*

Dalam penelitian, data memegang peranan yang sangat penting. Data yang sudah seimbang akan mempermudah pemrosesan lebih lanjut dalam penggunaan *machine learning*. Sebaliknya data yang tidak seimbang *(imbalance dataset)* akan mengakibatkan kesalahan dalam melakukan prediksi pada data baru yang diimplementasikan pada model yang dilatih. Data tidak seimbang dapat diakibatkan oleh beberapa faktor antara lain adanya *Biased Sampling* yang mana ketika pengambilan data terdapat bias atau memberatkan sebelah pihak serta adanya masalah pengukuran pada saat pengambilan data (*measurements Errors*) (Khalilia et al., 2011).

Kepedulian terhadap data yang tidak seimbang sangat dibutuhkan karena Ketika pengukuran performa menggunakan metrik akan menghasilkan interpretasi yang menyesatkan. Seperti contohnya apabila menggunakan metrik akurasi untuk mengukur performa model yang merupakan metrik yang meringkas performa model klasifikasi secara keseluruhan dengan cara menghitung total prediksi benar dibagi oleh total semua prediksi. Metrik akurasi bisa menyesatkan apabila datanya tidak seimbang sehingga disebut "*The Accuracy Paradox*" dimana model memiliki akurasi yang tinggi tetapi tidak dapat digunakan untuk prediksi secara nyata (Shelke et al., 2017).

Imbalanced data sendiri mengakibatkan adanya *imbalanced classification* yaitu suatu masalah klasifikasi dimana distribusi kelas target memiliki rasio berbeda jauh. Kelas yang mengambil proporsi terbesar pada data tersebut disebut

dengan kelas mayoritas sedangkan kelas yang mengambil proporsi terkecil disebut dengan kelas minoritas (Johnson & Khoshgoftaar, 2019).

2.3.7. Teknik Resampling

Teknik *Resampling* merupakan suatu teknik atau metode untuk membuat sampel baru dari sampel atau populasi yang sudah ada pada data. Metode ini bisa dibagi jadi dua kategori yaitu menghapus sampel dari kelas mayoritas sehingga rasio sama (*Undersampling*) dan yang kedua adalah menambah sampel ke kelas minoritas sehingga rasio sama (*Oversampling*) (Shelke et al., 2017).

Teknik *Undersampling* terdiri dari:

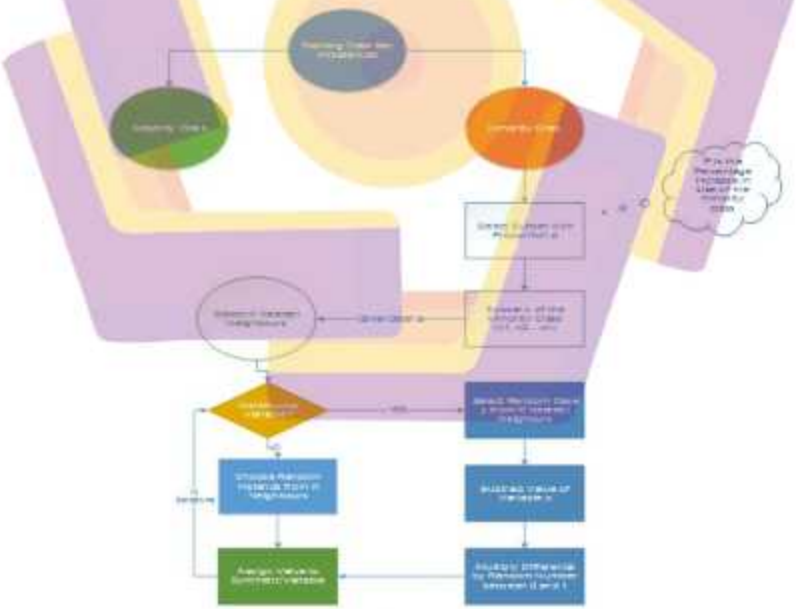
- *Random Undersampling*, dilakukan dengan menghapus sampel kelas mayoritas secara acak;
- *Prototype Generation*, dilakukan dengan menghapus sampel dan menambah sampel berdasarkan metode *clustering*;
- *Near Miss*, dilakukan dengan cara menghapus sampel kelas mayoritas yang dekat dengan kumpulan data kelas minoritas;
- *Tomek's Link*, dilakukan dengan menghapus sampel kelas mayoritas yang berpasangan dengan datapoin kelas minoritas.

Adapun teknik *Oversampling* sebagai berikut:

- *Random Oversampling*, dengan cara menambah sampel kelas minoritas secara acak dari sampel yang sudah ada.
- *SMOTe (Synthetic Minority Oversampling Technique)*, dilakukan dengan menambah sampel kelas minoritas dengan cara mensintesis data baru berdasarkan metode k-*Nearest Neighbour*.

2.3.8. *Synthetic Minority Oversampling Technique* (SMOTe)

SMOTe (*Synthetic Minority Oversampling Technique*) merupakan pendekatan *oversampling* yang secara sintetis menghasilkan *instance* dengan memilih *instance* secara acak dari kelas minoritas dan menggunakan metode interpolasi untuk menghasilkan *instance* antara titik yang dipilih dan *instance* yang berdekatan (Shelke et al., 2017). Teknik SMOTe akan mengambil sampel data menggunakan *up-sampling* dan *down-sampling*, tergantung pada kelasnya, yang memiliki tiga parameter operasional yaitu jumlah *up-sampling*, jumlah *down-sampling*, dan jumlah tetangga yang digunakan untuk memasukan kasus baru (Chawla et al., 2002).



Gambar 2.4. Proses SMOTE

Pada gambar 2.4. di atas dapat dilihat bahwa proses SMOTE berawal dari *training imbalance dataset* yang terbagi menjadi kelas mayoritas dan kelas minoritas. Kemudian dari kelas minoritas diambil subset data sebagai contoh dan kemudian *synthetic instance* baru yang sejenis sintetik baru. *Synthetic instance* ini kemudian ditambahkan ke dataset asli. Dataset baru yang terbentuk digunakan sebagai sampel untuk melatih model klasifikasi.

2.3.9. Pengukuran Kinerja Algoritma Klasifikasi

Setelah melakukan analisis pada metode klasifikasi dan didapatkan hasil prediksi, Langkah berikutnya melakukan kinerja algoritma untuk membandingkan nilai prediksi yang paling baik. Secara umum pengukuran kinerja klasifikasi dilakukan dengan membandingkan antara nilai prediksi algoritma klasifikasi dengan nilai target variable data *testing* sebagai data sebenarnya.

Dalam penelitian yang menggunakan *machine learning* umumnya menggunakan empat macam matriks untuk mengukur kinerja model, yaitu *accuracy, precision, recall* dan *f1 score* (Hutter & Kotthoff, 2019). *Imbalanced Dataset* adalah dataset yang memiliki contoh kelas negatif (kelas mayoritas) jauh lebih banyak daripada contoh kelas positif (kelas minoritas) (W. Wang & Zhang, 2011). Dalam mempelajari data yang tidak seimbang, akurasi klasifikasi secara keseluruhan seringkali bukan ukuran kinerja yang tepat (Thejas et al., 2022). Untuk data tidak seimbang, akurasi lebih didominasi oleh ketepatan pada data kelas minoritas, maka matriks yang tepat adalah AUC (*Area Under the ROC Curve*), *F-Measure*, *G-mean*, *Apparent Error Rate* (APER), *Total Accuracy Rate* (1-APER), dan akurasi kelas minoritas (W. Wang & Zhang, 2011). Evaluasi kinerja model

klasifikasi didasarkan pada pengujian objek yang diprediksi dengan benar dan salah, hitungan ini ditabulasikan *confusion matrix* (Hong & Gyu, 2021).

2.3.10 Evaluasi Kinerja

i. Confusion Matrix

Untuk mengukur kinerja klasifikasi digunakan *confusion matrix* yang memberikan keputusan yang diperoleh dalam pelatihan dan pengujian (Xu et al., 2020). *Confusion Matrix* adalah alat yang berguna untuk menganalisis seberapa baik *classifier* dapat mengenali tupel dari kelas yang berbeda dimana kelas yang diprediksi akan ditampilkan dibagian atas *matrix* dan kelas yang diobservasi ditampilkan dibagian kiri (Hong & Gyu, 2021). Berikut adalah tabel matriks *confusion* seperti Tabel 2.2 di bawah ini.

Tabel 2.2. *Confusion Matrix*

	Predicted Positive	*Predicted Negative*
Actual Positive	*TP (True*	***FN (False Negative)***
Actual Negative	***FP (False***	***TN (True Negative)***

dimana:

TP (*True Positive*) adalah jumlah observasi positivif yang tepat prediksi.

TN (*True Negative*) adalah jumlah observasi negatif yang tepat prediksi.

FP (*False Positive*) adalah jumlah observasi positif yang salah diprediksi sebagai negatif.

FN (*False Negative*) adalah jumlah observasi negatif yang salah diprediksi sebagai positif.

False Positive dikenal sebagai *error* tipe 1, terjadi ketika kasus yang seharusnya diklasifikasikan sebagai negatif diklasifikasikan positif sedangkan

false negative dikenal sebagai *error* tipe 2 terjadi jika kasus yang seharusnya diklasifikasikan sebagai positif diklasifikasikan negatif (Bramer et al., 2013).

Menurut (S. Wang & Yao, 2013), akurasi kelas minoritas dapat menggunakan matriks *True Positive Rate* atau *recall* (sensitivitas), *G-mean* dan AUC (*Area Under the ROC Curve*) merupakan evaluasi prediktor yang lebih komprehensif dalam konteks ketidakseimbangan. Dengan menggunakan matriks seperti *true negative rate* (*specificity*), *true positive rate* (*sensitivity*), *precision*, dan *F1-score* serta AUC (*Area Under the ROC Curve*) untuk dapat mengevaluasi kinerja *Machine learning* pada saat data yang tidak seimbang (Agarwal et al., 2021).

Presisi diperlukan karena metode klasifikasi cenderung baik dalam memprediksi kelas dengan data sampel yang lebih banyak namun buruk dalam memprediksi kelas dengan data sampel yang sedikit (Bramer et al., 2013). Nilai presisi merupakan akurasi dari prediksi positif seperti terlihat pada persamaan (1).

$$P = \frac{T_P}{T_P + F_P} \quad (1)$$

Sensitivitas mengukur proporsi *true positive* yang diidentifikasikan dengan benar sedangkan spesifitas mengukur proporsi *true negative* yang diidentifikasikan dengan benar (Bramer et al., 2013). *Recall* sendiri merupakan *positive instance* yang dideteksi secara benar oleh *classifier* yang terlihat pada persamaan (2).

$$R = \frac{T \quad P(}{T \quad P \quad +F \quad N} \tag{2}$$

Untuk meningkatkan *recall* atau *sensitivity* tanpa mempengaruhi *precision* adalah tujuan utama data *training* dari dataset *imbalanced*, namun tujuan dari *precision* dan *recall* sering bersebrangan karena ketika meningkatkan nilai *true positive* kelas minortias, jumlah *false positive* juga meningkat dan menyebabkan nilai *precision* berkurang (Miao & Zhu, 2021).

$$F1 - s = 2\,x\frac{P}{p} \quad \frac{xP}{+r} \tag{3}$$

Pada persamaan (3), *F1-score* merupakan metrik evaluasi yang menggabungkan angka *precision* dan *recall* sebab kedua nilainya bisa mempunyai bobot yang berbeda. Sehingga, *F1-score* merupakan rata-rata harmonis yang diperoleh dari hasil *precision* dan *recall*, rentang nilai dari 0 hingga 1.

ii. Area Under Curve (AUC)

Area Under Curve (AUC) merupakan metode umum yang digunakan untuk menghitung under *Receiver Operating Characteristic* (ROC) yang dibuat berdasarkan berdasarkan nilai yang diperoleh dari perhitungan menggunakan *confusion matrix*, yaitu antara *False Positive Rate* dengan *True Positive Rate* yang mana kinerja klasifikasi dikatakan bagus apabila mendekati titik (0,1). AUC dapat diartikan sebagai probabilitas, jika memilih satu contoh positif dan negatif secara acak, metode klasifikasi akan memberikan nilai lebih tinggi pada contoh positif daripada contoh negatif. AUC adalah luas area di bawah kurva ROC yang

merupakan integral dari fungsi ROC (Miao & Zhu, 2021). *AUC* merupakan ukuran numerik untuk membedakan kinerja model dan menujukkan seberapa sukses dan benar peringkat model dengan memisahkan pengamatan positif dan negatif (Pérez et al., 2020). AUC merangkum informasi kinerja pengklasifikasi ke dalam satu angka yang memperoleh perbandingan model ketika tidak ada kurva ROC yang mendominasi (Stern, 2021). AUC merupakan cara yang baik untuk mendapatkan nilai kinerja pengklasifikasi secara umum dan untuk membandingkannya dengan pengklasifikasian yang lain (Orynbassar et al., 2022). AUC adalah ukuran kinerja yang populer dalam ketidakseimbangan kelas dimana jika nilai AUC tinggi menunjukkan kinerja yang lebih baik sehingga untuk memilih model terbaik dengan cara menganalisa nilai AUC (Huang et al., 2021).

Untuk mencari nilai AUC menggunakan formula yang didasarkan pada confusion matrix seperti persamaan 4 dibawah ini:

$$A = \frac{1}{2}\left(\left(\frac{T}{T}\frac{P}{P}\right)\left(\frac{P}{+F}\frac{}{N}\right) + \left(\frac{T}{T}\frac{N}{P}\right)\left(\frac{N}{+F}\frac{}{N}\right)\right) \quad (4)$$

Nilai AUC akan selalu berada pada range 0-1, karena bagian dari luas persegi satuan dengan sumbu x dan sumbu y memiliki nilai dari 0 sampai 1. Nilai diatas 0,5 dikatakan nilai yang menarik karena prediksi acak menghasilkan garis diagonal antara (0,0) dan (1,1) yang memiliki luas 0,5. Kualitas klasifikasi keakuratan dari tes diagnostik menggunakan nilai AUC ditunjukkan pada tabel 2.2 (Gorunescu, 2010).

Tabel 2.3. Keakuratan hasil klasifikasi berdasarkan nilai AUC

Sumber: (Gorunescu, 2010)

Nilai AUC	Kategori
0.90-1.00	Sangat Baik
0.80-0.90	Baik
0.70-0.80	Cukup Baik
0.60-0.70	Kurang Baik
0.50-0.60	Buruk

Dari Tabel 2.3 dapat dilihat berdasarkan Gorunescu, keakuratan hasil klasifikasi berdasarkan nilai AUC dengan rentang 0,50-0,60 diklasifikasikan buruk, skala 0,60-0,70 diklasifikasikan dengan kurang baik, rentang 0,70-0,80 diklasifikasikan cukup baik, rentang 0,80-0,90 diklasifikasikan baik serta rentan 0,90-1,00 yang dikategorikan sangata baik.

BAB III

METODE PENELITIAN

3.1. Jenis, Sifat, dan Pendekatan Penelitian

Jenis penelitian ini adalah penelitian terapan ekperimen, Penelitian ini dilakukan dengan cara meneliti keberadaan *Lumpy Skin Disease* menggunakan *dataset* LSD yang dipublikasikan melalui Mendeley Data. Penelitian ini bersifat ilmiah yaitu sistematis, empiris, dan rasional. Penelitian dilakukan dengan sistematis yang berarti bahwa kegiatan penelitian memiliki langkah-langkah yang urut dan sistematis. Empiris kegiatan penelitian dapat diamati oleh indera manusia dan rasional kegiatan penelitian dapat terjangkau oleh nalar manusia (Anggara & Abdillah, 2019) .

Pendekatan penelitian yang digunakan adalah metode campuran dari penelitian deskriptif dan penelitian kuantiatif. Penelitian deskriptif digunakan untuk untuk mendeskriptifkan data-data yang dikumpulkan agar bisa memecahkan masalah penelitian. Sedangkan metode penelitian kuantitatif digunakan untuk meneliti pada populasi atau sampel tertentu, pengumpulan data menggunakan intrumen penelitian, analisis data bersifat statistik, dengan tujuan untuk menguji hipotesis yang telah ditetapkan (Pakpahan et al., 2022).

3.2. Metode Pengumpulan Data

Pengumpulan data dilakukan dengan mencari dataset dan mengumpulkan dokumen pendukung dari buku-buku dan sumber referensi lain yang berhubungan dengan *lumpy skin disease* untuk mendapatkan bahan yang relevan sebagai subjek penelitian. Observasi yang dilakukan dengan pengamatan langsung atau tidak langsung terhadap peristiwa yang diamati. Observasi untuk memperoleh data yang akan diolah merupakan dataset yang berasal dari subjek penelitian dan dilakukan secara tidak langsung.

Data yang digunakan pada penelitian ini merupakan jenis data sekunder. Data tersebut diperoleh dari Mendeley data *(doi: 10.17632/7pyhbzb2n9.1)* berupa dataset *Lumpy Skin Disease* yang berisi 24.804 baris data dengan dua puluh atribut data yang terbagai menjadi sepuluh atribut data dengan fitur-fitur meteorologi dan sepuluh atribut data dengan fitur-fitur geospasial seperti pada tabel 3.1 dan 3.2 berikut ini.

Tabel 3.1 Operasional Variabel Penelitian Fitur-Fitur Meteorologi

No.	Nama variabel	Definisi Operasional	Tipe Data
1	cld	*Monthly Cloud Cover* dalam persen	Float
2	dtr	*Diurnal Temperature Range* merupakan suhu harian dalam derajat celcius	Float
3	frs	*frost day frequency* merupakan jumlah frekuensi *frost day* dalam sebulan	Float
4	pet	*potential evapotranspiration* merupakan banyaknya penguapan dari suatu permukaan yang mana setiap komplkes perubahan air menjadi uap jika kelembaban mencapai 100% (evapotranspirasi potensial) dalam milimeter per hari	Float
5	pre	*precipitation* merupakan setiap produk dari kondensasi uap air di atmosfer dalam milimeter per bulan	Float

Tabel 3.1. (Lanjutan)

6	tmn	*daily mean temperature* merupakan suhu rata-rata harian dalam derajat *celcius*	Float
7	tmp	*temperature* merupakan suhu udara dalam derajat *celcius*	Float
8	tmx	*monthly average maximum and minimum temperature* merupakan suhu maksimum dan minimum rata-rata bulanan dalam derajat *celcius*	Float
9	vap	*vapor pressure* merupakan tekanan uap yang terjadi dalam *hectoposcal*	Float
10	wet	*wet day frequency* merupakan jumlah frekuensi *wet day* dalam hari	Float

Tabel 3.2 Operasional Variabel Penelitian Fitur-Fitur Geospasial

No.	Nama variabel	Definisi Operasional	Tipe Data
1	x	*lotitude* sumbu x koordinat spasial	Float
2	y	*longitude* sumbu y koordinat spasial	Float
3	Region	wilayah benua terjadinya wabah	String
4	country	negara terjadinya wabah	String
5	Reporting date	tanggal pelaporan terjadinya wabah	Date
6	elevation	ketinggian lokasi geografis dalam meter	Integer
7	dominant_landcover	tutupan lahan yang dominan	Integer
8	X5_Ct_2010_Da	*quick view file GIS* dari *dasymetric cattle*	image
9	X5_Bf_2010_Da	*quick view file GIS* dari *dasymetric buffalo*	image
10	lumpy	klasifikasi apakah terinfeksi LSD dengan kode: 1, tidak terinfeksi dengan kode: 0	Categorical

Tabel 3.1 berisi atribut dari dataset LSD yang berupa fitur-fitur meteorologi dan tabel 3.2 berisi atribut data dengan fitur-fitur geospasial yang akan digunakan dalam pemodelan prediksi keberadaan LSD menggunakan *Random Forest Classifier* dan teknik *resampling* SMOTe.

3.3. Metode Analisis Data

Pada penelitian ini metode analisis data yang digunakan adalah metode klasifikasi Random Forest dan teknik *resampling* SMOTe. Algoritma ini digunakan untuk mengklasifikasikan keberadaan virus LSD pada hewan ternak berdasarkan fitur meteorologi dan geospasial pada dataset LSD, dengan menggunakan bahasa *phyton* pada platform *Google Colabs*.

Metode klasifikasi yang digunakan dalam penelitian ini berupa *supervised learning* yaitu jenis *machine learning* di mana mesin dilatih menggunakan data pelatihan yang "diberi label" dengan baik, dan berdasarkan data tersebut, mesin memprediksi hasilnya. Data berlabel berarti beberapa data input sudah ditandai dengan output yang benar. Dalam *supervised learning*, data pelatihan yang diberikan ke mesin berfungsi sebagai pengawas yang mengajarkan mesin untuk memprediksi output dengan benar.

Supervised learning adalah proses memberikan data input serta data output yang benar ke model *machine learning* yang bertujuan untuk menemukan fungsi pemetaan untuk memetakan variabel masukan (x) dengan variabel keluaran (y).

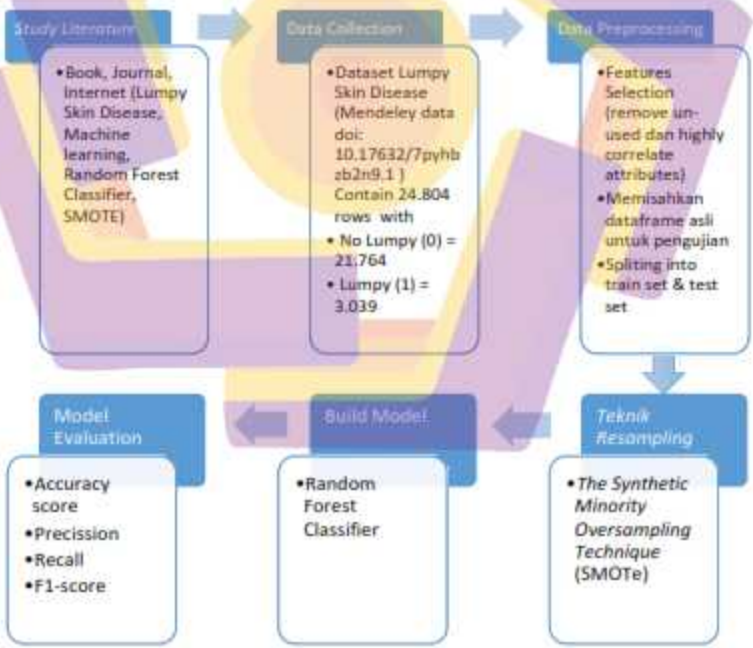
3.4. Alur Penelitian

Penelitian ini terbagi menjadi enam tahapan penelitian. Tahap pertama adalah studi pustaka yang mengambil literatur dari berbagai sumber Pustaka, kemudian dilanjutkan dengan pengambilan data. Tahap selanjutnya adalah pra-pemprosesan data yang terdiri dari seleksi fitur dan pembagian data pelatihan dan data validasi. Setelah itu dilakukan teknik resampling SMOTe. Pengujian terhadap

model yang telah dilatih dilakukan yang kemudian akan dilakukan evalusi terhadap kinerja model prediksi.

Langkah pertama dalam penelitian ini yaitu studi Pustaka yang bertujuan untuk mencari referensi mengenai apa itu *Lumpy Skin Disease (LSD)*, bagaimana penularannya serta kondisi apa saja yang membuat penyakit tersebut menginfeksi ke hewan ternak yang lainnya. Selain itu dilakukan juga kajian mengenai bagaimana cara menangani *dataset* yang tidak seimbang sebelum dilakukan pra-pemrosesan data lebih lanjut.

Adapun bagan alur penelitiannya seperti gambar 3.1 berikut ini.



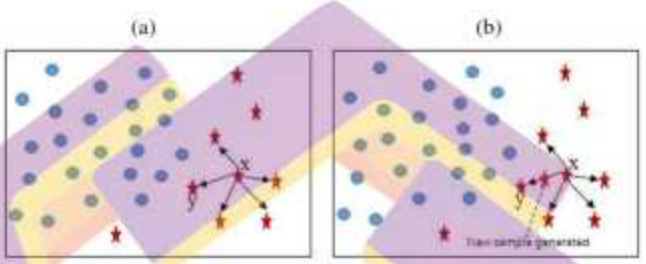
Gambar 3.1. Bagan Alur Penelitian

Pencarian referensi terkait algoritma yang akan digunakan untuk melakukan pemrosesan data serta melakukan prediksi terhadap *dataset* LSD, yaitu *Random forest Classifier*. Informasi yang didapat dari referensi buku, jurnal maupun sumber internet berguna dalam pembangunan model serta mengarahkan langkah-langkah selanjutnya dari teori yang telah diperoleh. Sumber data yang telah tercatat sebelumnya digunakan sebagai rujukan yang diperoleh dengan menerapkan metode penelitian.

Data yang digunakan untuk penelitian diambil dari Mendeley data berupa *dataset lumpy skin disease* yang terdiri dari 20 atribut data dan 24.803 baris data. Seleksi fitur adalah proses pengurangan jumlah variabel input ketika mengembangkan model prediktif. Diinginkan untuk mengurangi jumlah variabel input untuk mengurangi biaya komputasi pemodelan dan, dalam beberapa kasus, untuk meningkatkan kinerja model. Metode pemilihan fitur berbasis statistik melibatkan evaluasi hubungan antara setiap variabel input dan variabel target menggunakan statistik dan memilih variabel input yang memiliki hubungan terkuat dengan variabel target.

Dataset LSD terdiri dari data ternak yang terinfeksi LSD hanya 3.039 data atau 12,25% dari keseluruhan total data. Sehingga sebesar 21.764 data atau 87,75% merupakan data ternak yang tidak terinfeksi sehingga dataset LSD yang akan digunakan untuk pemodelan merupakan dataset yang tidak seimbang yang harus dilakukan *resampling* terlebih dahulu agar hasil klasifikasi tidak bias. Resampling terhadap dataset LSD akan menggunakan teknik SMOTe (*Synthetic Minority Oversampling Technique*) yaitu akan mensintesis jumlah data terinfeksi LSD

sebanyak data yang tidak terinfeksi LSD. Setelah dilakukan resampling dengan SMOTe maka data yang terinfeksi LSD dan yang tidak terinfeksi masing-masing berjumlah 21.764 data dengan total data keseluruhan menjadi 43.528 data. Generasi *sampel* SMOTe diilustrasikan pada Gambar 3.2 berikut ini.



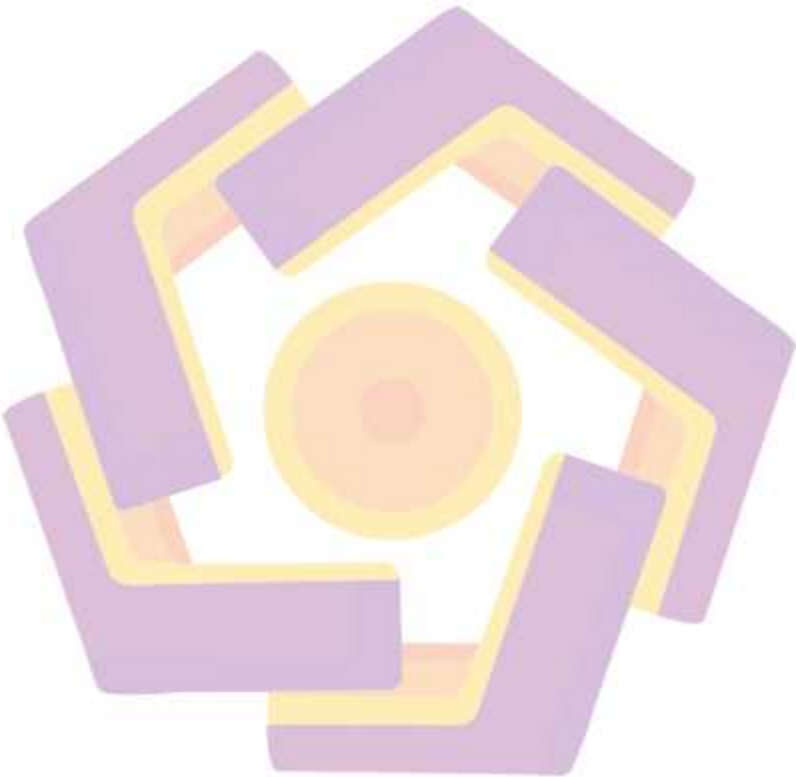
Gambar 3.2. Pembuatan sampel baru di SMOTe.

Pada gambar 3.2. bagian (a) menunjukan k-Neigbour terdekat yang sama sampel dari sampel utama x (K=5, dalam contoh ini) sedangkan pada bagian (b) terlihat sampel baru yang dihasilkan.

Sebelum melakukan analisis klasifikasi dilakukan pembagian data menjadi dua bagian, yaitu *data training* dan *data testing*. Data *training* berguna melatih algoritma untuk pembentukan sebuah model, dan data *testing* digunakan untuk mengukur sejauh mana tingkat keakuratan dan performa yang didapatkan dari data training. Data *training* dan data *testing* dibagi dengan proporsi 80% untuk data *training* dan 20% dari data *testing* dari total *dataset* menggunakan *train_test_split* dari *scikit learn library*.

Setelah melakukan pembagian *data training* dan *data testing* tahapan selanjutnya yaitu melakukan analisis klasifikasi *Random Forest*. Kemudian evaluasi model dilakukan dengan melakukan prediksi terhadap data testing

menggunakan model pohon keputusan pada Random Forest yang terbentuk. Dengan mengevaluasi model digunakan nilai *precision, recall,* dan *F1-Score* yang menggambarkan hubungan antara data testing dan data prediksi.



BAB IV

HASIL PENELITIAN DAN PEMBAHASAN

Berdasarkan kajian teori dan hasil penelitian sebelumnya, maka pada bab ini peneliti akan menjelaskan hasil-hasil yang telah didapatkan dari hasil komputasi klasifikasi menggunakan metode Random Forest dan teknik resampling SMOTe. Pembahasan pada bab ini meliputi analisis deskriptif, *data preprocessing*, penentuan data pelatihan dan validasi, penentuan jumlah *N-tree* terbaik, klasifikasi menggunakan metode Random Forest dan teknik *resampling* serta mengevaluasi hasil eksperimen atas metode klasifikasi yang digunakan.

4.1. Langkah-Langkah Penelitian

Adapun Langkah-langkah yang dilakukan dalam penelitian ini meliputi analisis deskriptif, pra premrosesan data yang berupa seleksi fitur dan pembagian data pelatihan dan validasi, penerapan teknik *resampling* terhadap *dataset* yang ada, kemudian melakukan klasifikasi menggunakan data asli maupun data yang telah diresampling serta melakukan evaluasi terhadap kinerja model.

4.1.1 Analisis Deskriptif

Salah satu tahapan yang dilakukan sebelum melakukan analisis data adalah analisis deskriptif yang digunakan untuk menggambarkan atau mendeskripsikan masing-masing variabel yang digunakan dalam penelitian. Pertama peneliti akan melihat tipe data setiap variabel yang digunakan dalam penelitian menggunakan fungsi *.info* dan *.describe* pada phyton dengan hasil sebagai berikut.

```
df.info()
df.describe()
```

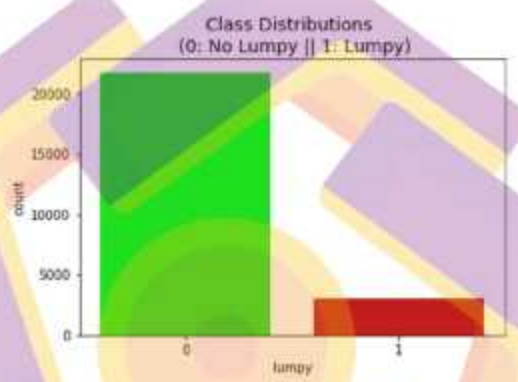
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 24803 entries, 0 to 24802
Data columns (total 20 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   x                      24803 non-null  float64
 1   y                      24803 non-null  float64
 2   region                 3039 non-null   object
 3   country                3039 non-null   object
 4   reportingDate          3039 non-null   object
 5   cld                    24803 non-null  float64
 6   dtr                    24803 non-null  float64
 7   frs                    24803 non-null  float64
 8   pet                    24803 non-null  float64
 9   pre                    24803 non-null  float64
 10  tmn                    24803 non-null  float64
 11  tmp                    24803 non-null  float64
 12  tmx                    24803 non-null  float64
 13  vap                    24803 non-null  float64
 14  wet                    24803 non-null  float64
 15  elevation              24803 non-null  int64
 16  dominant_land_cover    24803 non-null  int64
 17  X5_Ct_2010_Da          24803 non-null  float64
 18  X5_Bf_2010_Da          24803 non-null  float64
 19  lumpy                  24803 non-null  int64
dtypes: float64(14), int64(3), object(3)
```

Gambar 4.1. Tipe Data Variabel-Variabel Penelitian

Pada gambar 4.1 terdapat 24.803 data observasi dengan jumlah kolom sebanyak 20 variabel. Variabel *lumpy* merupakan variabel *dependent* (terikat) yaitu kasus terinfeksi *lumpy*, dengan dua kategori yaitu terinfeksi *lumpy* (1) dan tidak terinfeksi *lumpy* (0). Selanjutnya variabel *independent* yang terdiri dari *x, y, region, country, reportingDate, cld, dtr, frs, pet, pre, tmn, tmp, tmx, vap, wet, elevation, dominant_land_cover, X5_Ct_2010_Da,* dan *X5_Bf_2010_Da.* Variabel *x, y, cld, dtr, frs, pet, pre, tmn, tmp, tmx, vap, wet, X5_Ct_2010_Da,* dan *X5_Bf_2010_Da* bertipe data float (tipe data yang dipergunakan untuk variabel-variabel yang memiliki nilai pecahan/desimal). Selanjutnya variabel *region, country, dan reportingDate* memiliki tipe data *object* (Pandas menggunakan objek *ndarray*, yang menyimpan pointer ke objek; karena ini *dtype ndarray* jenis ini adalah objek).

Kemudian variabel *elevation* dan *dominant_land_cover* bertipe data int64 (tipe data yang berisi kumpulan bilangan bulat dengan rentang nilai yang ditampung 2 pangkat 64).

Selanjutnya peneliti akan melihat gambaran umum distribusi kelas pada variabel *dependent lumpy*.



Gambar 4.2. Distribusi Kelas

Distribusi kelas pada dataset LSD pada Gambar 4.2. tersebut dapat terlihat bahwa *dataset* tidak seimbang dimana data dengan kasus terinfeksi *lumpy* hanya sebesar 12,25% dari keseluruhan data dibandingkan dengan data pada kasus tidak terinfeksi *lumpy* yang memiliki prosentase 87,75% dimana sebanyak 3.039 sapi terinfeksi LSD dan 21.764 sapi tidak terinfeksi LSD. Berdasarkan hal tersebut maka dataset yang ada diperlukan proses *resampling* untuk menyeimbangkan datanya agar hasil dari klasifikasi yang dilakukan tidak bias. Peneliti ingin mengetahui apa saja faktor-faktor yang mempengaruhi sapi terinfeksi LSD dan meminimalisir terjadinya infeksi LSD.

4.1.2 Pra-pemrosesan Data

Dalam melakukan pemodelan data terhadap prediksi terhadap keberadaan LSD berdasar fitur-fitur metereologi dan geospasial ini, akan dilakukan pra-pemrosesan data terlebih dahulu agar data tersebut sesuai untuk membangun dan melatih model *Machine Learning*. Sebelum melakukan pemodelan data, akan dilakukan seleksi fitur yang berupa penghapusan atribut yang tidak digunakan dalam proses klasifikasi. Adapun penghapusan atribut tersebut mengacu pada penelitian sebelumnya menggunakan *dataset* yang sama seperti atribut *'region'*, *'country'*, *'reportingDate'*, *'X5_Ct_2010_Da'*, dan *'X5_Bf_2010_Da'* (Safavi, 2022).

Penghapusan atribut ini menggunakan fungsi .drop dari *library pandas* di phyton seperti berikut ini.

```
df=df.drop(['region', 'country', 'reportingDate', 'X5_Ct_2
010_Da', 'X5_Bf_2010_Da'], axis=1)
df=df.dropna()
rows = df.shape[0]
df = df.drop(np.array(range(24803,rows)), axis=0)
df.info()
df.describe()
df.sample(frac=1.).reset_index(drop=True,inplace=True)
df.head()
```

Fungsi *.drop* dari *library* pandas pada phyton untuk menghapus atribut pada dataframe LSD yang terdiri dari kolom *'region'*, *'country'*, *'reportingDate'*, *'X5_Ct_2010_Da'*, dan *'X5_Bf_2010_Da'* berserta isi datanya.

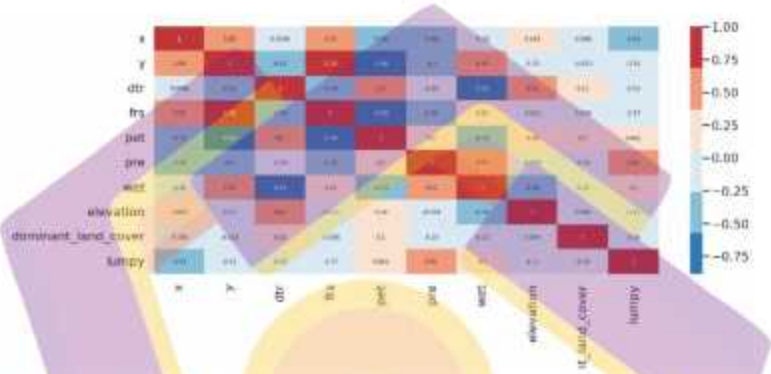
Selain itu dilakukan pengecekan korelasi antar atribut menggunakan matrik korelasi untuk mengurangi atribut yang saling berkorelasi tinggi (Ramsay et al., 1984). Matriks korelasi adalah alat penting dari analisis data eksplorasi (Kahl & Günther, 2008). *Correlation heatmap* berisi informasi yang sama dengan cara yang

menarik secara visual yang menunjukkan sekilas variabel mana yang berkorelasi, sejauh mana, ke arah mana, dan mengingatkan pada potensi masalah multikolinearitas yang ditampilkan dengan menggunakan *Seaborn* pada *library Matplotlib* untuk visualisasi data di Phyton. Adapun matrik korelasi tersebut divisualisasikan dengan *Correlation Heatmap* yang dapat dilihat pada gambar di bawah ini.

Gambar 4.3. *Correlation heatmap dataset* LSD

Berdasar gambar 4.3. terdapat beberapa fitur yang memiliki korelasi tinggi. Atribut *tmn* dan *tmx* memiliki korelasi 0,99, *tmn* dan *tmp* memiliki korelasi 1, *tmp* dan *tmx* berkorelasi 1 yang pada dasarnya memiliki arti yang sama. Atribut *tmn*, *tmp* dan *tmx* memiliki korelasi yang sangat tinggi dengan atribut *vap* (0,86, 0,86 dan 0,85). Dua variabel berkorelasi tinggi lainnya adalah *tmn* dan *vap*, yang mana *vap* lebih berkorelasi dengan target (0,17). Atribut *cld* dengan *wet* berkorelasi 0,75 serta atribut *pet* dengan *vap* memiliki korelasi 0,85. Dengan mempertimbangkan korelasi antar atribut serta korelasi atribut dengan target maka akan dilakukan

penghapusan pada atribut *tmp*, *tmx*, *tmn*, *cld*, dan vap. Atribut *frs* dan *y* tidak dilakukan penghapusan karena atribut *y* merupakan atribut sama sekali berbeda. Hasil dari penghapusan fitur yang saling berkorelasi tinggi tersebut dapat dilihat pada gambar 4.5. di bawah ini.



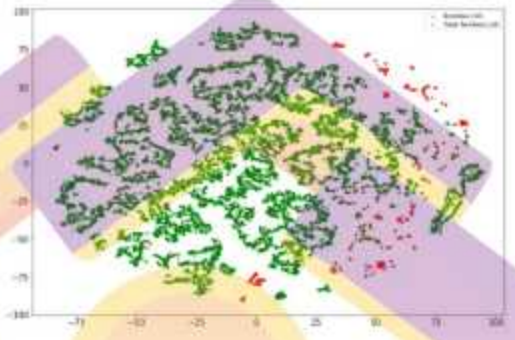
Gambar 4.4. *Correlation heatmap* setelah penghapusan fitur yang berkorelasi tinggi

Dari gambar 4.4. di atas dapat dilihat bahwa fitur yang akan digunakan untuk pemodelan data sebanyak sepuluh fitur yang terdiri *dari x, y, dtr, frs, pet, pre, wet, elevation, dominant_land_cover* dan *lumpy* dari total awal sebanyak 20 fitur.

4.1.3 *Resampling*

Teknik *resampling* terhadap *imbalanced dataset* LSD dengan menggunakan *SMOTe*. Hal ini dilakukan karena adanya ketidakseimbangan jumlah data antara yang terinfeksi dan yang tidak sehingga perlu dilakukan resampling terhadap dataset tersebut agar menjadi seimbang dan klasifikasi yang dihasilkan tidak bias. *Imbalanced data* sendiri mengakibatkan adanya *imbalanced classification* yaitu suatu masalah klasifikasi dimana distribusi kelas target memiliki rasio berbeda jauh.

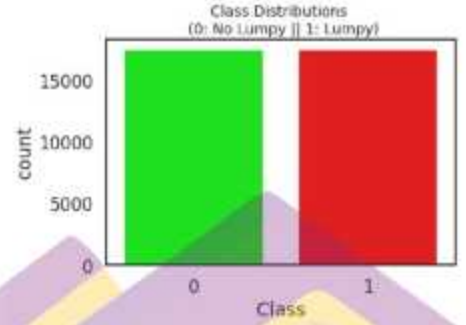
Kelas yang mengambil proporsi terbesar pada data tersebut disebut dengan kelas mayoritas sedangkan kelas yang mengambil proporsi terkecil disebut dengan kelas minoritas (Johnson & Khoshgoftaar, 2019). Adapun sebaran kelas-nya dapat dilihat seperti pada gambar 4.5. di bawah ini.



Gambar 4.5. Visualisasi hasil sebaran data sebelum dilakukan SMOTe pada *dataset* LSD

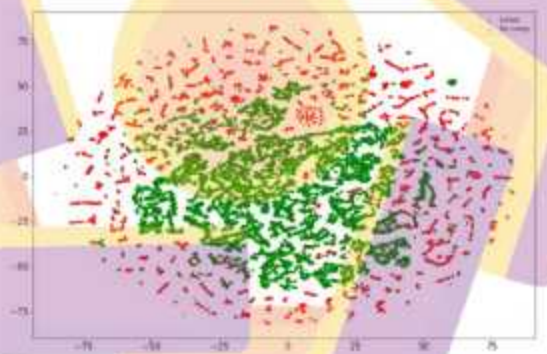
Pada gambar 4.5. terlihat bahwasanya sebaran data dalam dataset tidak seimbang dimana data yang berwarna merah (terinfeksi *lumpy*: 1) jauh lebih sedikit dari data berwarna hijau (yang tidak terinfeksi *lumpy*: 0). Oleh karena itu resampling terhadap *dataset LSD* diperlukan untuk menghindari adanya hasil dari model yang bias nantinya.

Teknik *oversampling SMOTe* menggunakan *fit_resample* menghasilkan *resampled dataset shape counter* ({1: 21.764, 0: 21.764}) yang berarti bahwa data terinfeksi *lumpy* (1) dan data tidak terinfeksi *lumpy* (0) masing-masing memiliki jumlah data yang sama yaitu 21.764 seperti terlihat pada Gambar 4.6. di bawah ini.



Gambar 4.6. Distribusi Kelas Setelah SMOTe

Visualisasi sebaran data antara kelas terinfeksi *lumpy* dan kelas yang tidak terinfeksi *lumpy* dapat dilihat pada Gambar 4.7. berikut ini.



Gambar 4.7. Visualisasi hasil sebaran data sesudah dilakukan SMOTe pada *dataset* LSD

Pada gambar 4.7. diatas merupakan visualisasi sebaran data sesudah dilakukan *resampling* dengan SMOTe. Terlihat perbedaan yang cukup jelas bahwa setelah dilakukan proses *resampling* data berwarna merah (yang terinfeksi *lumpy)* maupun data berwarna hijau (yang tidak terinfeksi *lumpy)* menjadi berimbang.

4.1.4 Penentuan Data Training dan Data Testing

Langkah selanjutnya yaitu melakukan pembagian *data training* dan *data testing* pada *dataset* LSD. *Data Training* sangat diperlukan dalam pemodelan *machine learning* dimana *data training* tersebut digunakan untuk melatih kinerja model. Jumlah data yang digunakan dalam penelitian ini berjumlah 24.803 data, dengan keterangan sebayak 3.039 sapi terinfeksi LSD dan 21.764 sapi tidak terinfeksi LSD. Data *training* dan data *testing* dibagi dengan proporsi 80% untuk data *training* dan 20% dari data *testing* dari total *dataset* menggunakan *train_test_split* dari *scikit learn library*. Pembagian data training dan data testing yang dilakukan pada *dataset LSD* asli seperti pada tabel 4.1 di bawah ini.

Tabel 4.1 *Data Training* dan *Data Testing* dari *Dataset LSD* Asli

	Train	*Test*	Total
Lumpy	2.431	608	3.039
No Lumpy	*17.411*	*4.353*	*21.764*
Total	19.842	4.961	24.803

Penentuan jumlah *data training* dan *data testing* pada Tabel 4.1 digunakan untuk pemodelan *random forest* menggunakan *dataset* LSD yang asli. Perolehan data training yang digunakan pada pelatihan model *random forest* sebanyak 2.431 data untuk kasus terinfeksi *lumpy* dan sebanyak 17.411 data untuk yang tidak terinfeksi *lumpy* dengan total keseluruhan data pelatihan sebanyak 19.842 buah data. Sedangkan untuk *data testing*-nya sebesar 20% dari keseluruhan *dataset*, didapatkan jumlah data yang digunakan untuk *testing* pada kasus terinfeksi *lumpy* sebanyak 608 buah data dan yang tidak terinfeksi *lumpy* sebanyak 4.535 buah data dengan total keseluruhan data *testing berjumlah* 4.961 buah data.

Karena adanya ketidakseimbangan kelas data yang sangat mencolok, maka pemodelan *random forest* juga menggunakan skenario kedua dengan penentuan jumlah *data training* dan *data testing* setelah dilakukan proses *resampling* menggunakan SMOTe. Adapun pembagian *data training* dan *data testing* yang dilakukan pada *dataset LSD* yang telah dilakukan teknik resampling SMOTe seperti pada tabel 4.2 di bawah ini.

Tabel 4.2 *Data Training* dan *Data Testing* setelah *Oversampling*

	Train	Test	Total
Lumpy	17.411	608	18.019
No Lumpy	17.411	4.353	21.764
Total	34.822	4.961	39.783

Tabel 4.2 menunjukan proporsi antara *data training* dan *data testing* setelah dilakukan proses *resampling* menggunakan SMOTe dengan pembanding yang sama sebesar 80:20 untuk *data training:data testing*. Data training pada kasus terinfeksi lumpy dan tidak terinfeksi lumpy mendapatkan jumlah data yang sama yaitu 17.411 karena pada proses resampling menggunakan SMOTe ini data antara kasus yang terinfeksi lumpy dan yang tidak terinfeksi disamakan. Sehingga total data training-nya menjadi 34.822 data. Sedangkan data *testing* memiliki total yang sama dengan *data testing* yang diambil dari *dataset original*, dikarenakan untuk keperluan *testing* memang menggunakan *dataset* asli bukan data yang telah disintesis. Jumlah *data testing* untuk kasus terinfeksi *lumpy* sebanyak 608 data dan data yang tidak terinfeksi yang digunakan untuk *testing* sebanyak 4.353 data dengan total keseluruhan 4.961 data.

4.1.5 Penentuan Jumlah Pohon Terbaik (*NTree*)

Secara umum, dalam *machine learning* dan ilmu data, sangat penting untuk membuat sistem tepercaya yang akan bekerja dengan baik dengan data baru yang tidak terlihat. Secara keseluruhan, ada banyak pendekatan dan metode yang berbeda untuk mencapai generalisasi ini. *Out-of-bag Error* adalah salah satu metode ini untuk memvalidasi model *machine learning*.

Klasifikasi Random Forest dilatih menggunakan agregasi *bootstrap*, di mana setiap *tree* baru cocok dari sampel bootstrap dari pengamatan pelatihan. *Out-of-bag Error (OOB)* adalah kesalahan rata-rata untuk masing-masing dihitung menggunakan prediksi dari *tree* yang tidak mengandung sampel *bootstrap* masing-masing. Hal ini memungkinkan klasifikasi Random Forest untuk cocok dan divalidasi saat sedang dilatih (Ibrahim et al., 2009). *OOB Error* dapat diukur pada penambahan setiap *tree* baru selama pelatihan. Plot yang dihasilkan memungkinkan untuk memperkirakan nilai n_estimator yang sesuai di mana kesalahan stabil.

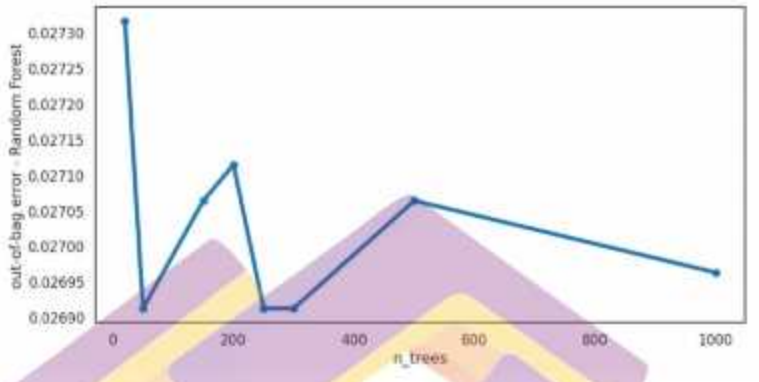
Fitur lain yang berguna dari Random Forest adalah konsep tingkat kesalahan *out-of-bag* (OOB). Karena hanya dua pertiga dari data yang digunakan untuk melatih setiap pohon saat membangun hutan, sepertiga dari data yang tidak terlihat dapat digunakan yang bisa menguntungkan hasil metrik akurasi tanpa kompleksitas komputasi yang tinggi seperti yang terjadi pada validasi silang.

Source Code Penentuan OOB Algoritma Random Forest

```
RF = RandomForestClassifier(oob_score=True,
                            random_state=42,
                            warm_start=True,
                            n_jobs=-1)

oob_list = list()
for n_trees in [20, 50, 150, 200, 250, 300, 500, 1000]:
    RF.set_params(n_estimators=n_trees)
    RF.fit(X_train, y_train)
    oob_error = 1 - RF.oob_score_
    oob_list.append(pd.Series({'n_trees': n_trees, 'oob': oo
b_error}))

rf_oob_df = pd.concat(oob_list, axis=1).T.set_index('n_trees
')
rf_oob_df
```

Seperti diuraikan di atas, saat menghitung OOB, dua parameter harus diubah. Dengan memanfaatkan *for-loop* di banyak ukuran hutan, tingkat kesalahan OOB dapat dihitung dan digunakan untuk menilai berapa banyak pohon yang sesuai untuk model yang dibangun.

Saat menghitung nilai oob, pengaturan *bootstrap=True* akan menghasilkan kesalahan, tetapi diperlukan dalam perhitungan *oob_score*. Nilai n_trees yang dimasukan dalam perhitungan oob error adalah 20, 50, 150, 200, 250, 300, 500, 1000. Hasil yang didapatkan dari perhitungan tersebut didapatkan nilai minimal kesalahan to witness pada n_trees adalah 250 dengan skor OOB 0,026913 yang merupakan skor OOB paling rendah diantara n_trees lainnya. Adapun Visualisasi skor OOB tersebut terlihat pada gambar 4.8 di bawah ini.



Gambar 4.8. Visualisasi *Out of Bag Error* pada Random Forest

Pada gambar 4.8. dapat dilihat bahwa minimum *error to witness* sebanyak 250 *tree* sehingga pemodelan pada klasifikasi RandomForest ini akan menggunakan n_estimator sebanyak 250.

n_trees	oob
20.0	0.027316
50.0	0.026913
150.0	0.027064
200.0	0.027114

n_trees	oob
250.0	0.026913
300.0	0.026913
500.0	0.027064
1000.0	0.026963

Gambar 4.9. Skor *Out of Bag Error* pada Random Forest

Dari gambar 4.9. dapat dilihat bahwa nilai OOB terendah dan stabil ada pada n_trees 250 dengan nilai 0,026913 sedangkan nilai OOB tertinggi pada n_trees 20 dengan nilai 0,027316

4.1.6 Klasifikasi Model

Proses klasifikasi model yang dilakukan menggunakan dua skenario yaitu klasifikasi RandomForest menggunakan data asli dan klasifikasi menggunakan data menggunakan data yang telah dilakukan *oversampling* terlebih dahulu yang dioptimasi dengan TPOT.

Dalam proses pelatihan metode Random Forest dibuat menjadi 250 iterasi berdasarkan dari penentuan *N-tree* terbaik menggunakan *OOB-Error* agar mendapatkan hasil yang lebih optimal. Untuk mendapatkan model Random Forest terbaik, hasil iterasi harus optimal karena hasil iterasi adalah jumlah pohon yang digunakan dalam *learning*.

Berikut *source code* penerapan model Random Forest terhadap data asli.

```
def fit_and_print(model):
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    print("Confusion Matrix: \n", confusion_matrix(y_test,
y_pred))
    print("Classification Report: \n", classification_repor
t(y_test, y_pred))
    print("Accuracy Score: ", accuracy_score(y_test, y_pred
))
    print("Recall score:", recall_score(y_test, y_pred))
    print("Precision score:", precision_score(y_test, y_pre
d))
    print("F1 score:", f1_score(y_test, y_pred))
    #AUC and KS
    print("AUC: ", roc_auc_score(y_test, y_pred))
    print("KS: ", ks_2samp(y_pred[y_test == 0], y_pred[y_te
st == 1]).statistic)
X = df[features].values
y = df['lumpy'].values
X_train, X_test, y_train, y_test = train_test_split(X, y, s
tratify=y, test_size=0.2, random_state=42)
print("Split between test and train!")
rf = RandomForestClassifier(n_estimators=250, max_features=
'auto', random_state=42)
fit_and_print(rf)
```

Tipe Random Forest yang digunakan adalah klasifikasi, dimana jumlah pohon yang digunakan sebanyak 250 yang diambil dari plot jumlah iterasi sebelumnya. Random_state diberikan nilai 42 yang digunakan untuk menginisialisasi *generator* nomor acak internal, yang akan memutuskan pemisahan data menjadi data pelatihan dan pengujian. Untuk *max_features* nya di set menjadi '*auto*'. Data yang digunakan untuk pelatihan berjumlah 19.858 data dan sebanyak 4.964 data digunakan sebagai data validasi.

Teknik *SMOTe* diterapkan untuk mengatasi ketidakseimbangan kelas data dengan cara meresampling eliminasi kelas minoritas (*lumpy*) secara acak sampai

jumlahnya sebanyak kelas mayoritas (*no lumpy*). Penggunaan data pelatihan dengan jumlah kelas data paling banyak yaitu kelas *no lumpy* dengan jumlah 17.411 data. Peneliti akan menyamaratakan kelas dengan masing-masing jumlah data sebanyak 17.411 data. Sebelum dilakukan teknik *SMOTe*, akan dilakukan pemisahan *data frame* asli yang akan digunakan pada saat pengujian model. Hal ini perlu dilakukan agar data yang digunakan pada pengujian merupakan data asli bukan data yang telah di sintesis agar mendapatkan hasil yang tidak bias.

Source Code Pemisahan *Dataframe* Asli:

```
from sklearn.model_selection import StratifiedKFold
sss = StratifiedKFold(n_splits=5, random_state=None, shuffle
=False)

for train_index, test_index in sss.split(X, y):
    print("Train:", train_index, "Test:", test_index)
    original_Xtrain, original_Xtest = X.iloc[train_index], X
.iloc[test_index]
    original_ytrain, original_ytest = y.iloc[train_index], y
.iloc[test_index]

original_Xtrain, original_Xtest, original_ytrain, original_y
test = train_test_split(X, y, test_size=0.2, random_state=42
)
original_Xtrain = original_Xtrain.values
original_Xtest = original_Xtest.values
original_ytrain = original_ytrain.values
original_ytest = original_ytest.values

train_unique_label, train_counts_label = np.unique(original_
ytrain, return_counts=True)
test_unique_label, test_counts_label = np.unique(original_yt
est, return_counts=True)
print('-' * 100)
print('Label Distributions: \n')
print(train_counts_label/ len(original_ytrain))
print(test_counts_label/ len(original_ytest))
```

Source code di atas menunjukan proses pemisahan dataframe asli yang akan digunakan dalam set pengujian. Data tersebut disimpan kedalam *variabel original_Xtrain, Original_Xtest, original_ytrain* dan *original_ytest*. Jadi variabel inilah yang nantinya akan digunakan pada set pengujian. Setelah *dataframe* original dipisahkan, maka teknik *oversampling* dapat dilaksanakan.

Source code teknik SMOTe.

```
sm = SMOTE(sampling_strategy='auto', random_state=42, k_ne
ighbors=5, n_jobs=-1)
X_res, y_res = sm.fit_resample(X_train, y_train)

np.bincount(y_res)
print('Resampled dataset shape {}'.format(Counter(y_res)))
```

Teknik SMOTe di atas menunjukan cara pengaplikasian teknik *oversampling* SMOTe pada phyton dimana konfigurasi yang digunakan untuk *sampling_strategy='auto', random_state=42, k_neighbors=5, n_jobs=-1*). Variabel hasil SMOTe disimpan dalam variabel X_res dan y_res. Hasil dari pengaplikasian SMOTe pada *dataframe* dapat terlihat seperti di bawah ini.

Hasil Pengaplikasian SMOTe pada *Dataframe*.

```
Resampled dataset shape Counter({1: 17411, 0: 17411})
```

Hasil Pengaplikasian SMOTe pada *Dataframe* di atas dimana untuk jumlah kelas *lumpy* dan *no lumpy* memiliki jumlah data pelatihan yang sama banyaknya yaitu masing-masing 17.411 buah data.

Klasifikasi yang dilakukan pada data yang telah di-*resampling* dengan SMOTe dilakukan dengan cara melakukan pelatihan model menggunakan data yang telah di-*resampling* dengan jumlah data yang sama untuk kelas *lumpy* dan *no lumpy* dan untuk data validasi yang digunakan berupa data asli sebesar 20% dari total data asli.

Source code prediksi menggunakan *Random Forest* dan SMOTe.

```
tpot_classifier.fit(X_res,y_res)
predpotsm = tpot.predict(original_Xtest)
print("Confusion Matrix: \n", confusion_matrix(original_ytes
t, predpotsm))
print("Classification Report: \n", classification_report(ori
ginal_ytest, predpotsm))
print('Accuracy score RF: {0:0.4f}'.format(accuracy_score(or
iginal_ytest, predpotsm)))
print("Recall Score RF:", recall_score(original_ytest, predp
otsm))
print("Precision score RF:", precision_score(original_ytest,
 predpotsm))
print("F1 score RF:", f1_score(original_ytest, predpotsm))
print("AUC: ", roc_auc_score(original_ytest, predpotsm))
print("KS: ", ks_2samp(predpotsm[original_ytest == 0], predp
otsm[original_ytest == 1]).statistic)
```

Source code di atas menunjukan proses prediksi yang dilakukan dengan data pelatihan berupa data yang telah memiliki jumlah yang sama untuk masing-masing kelas data dan dilakukan validasi menggunakan data asli.

4.2 Hasil Penelitian

Pada bagian ini penulis akan membahas mengenai hasil penelitian yang telah dilakukan mengenai prediksi *lumpy skin disease* menggunakan dua buah skenario pemodelan yaitu prediksi menggunakan klasifikasi *Random Forest* yang dilakukan

terhadap data asli serta klasifikasi algoritma *Random Forest* pada dataframe yang telah di *resampling* dengan teknik SMOTe.

4.2.1 Klasifikasi Random Forest

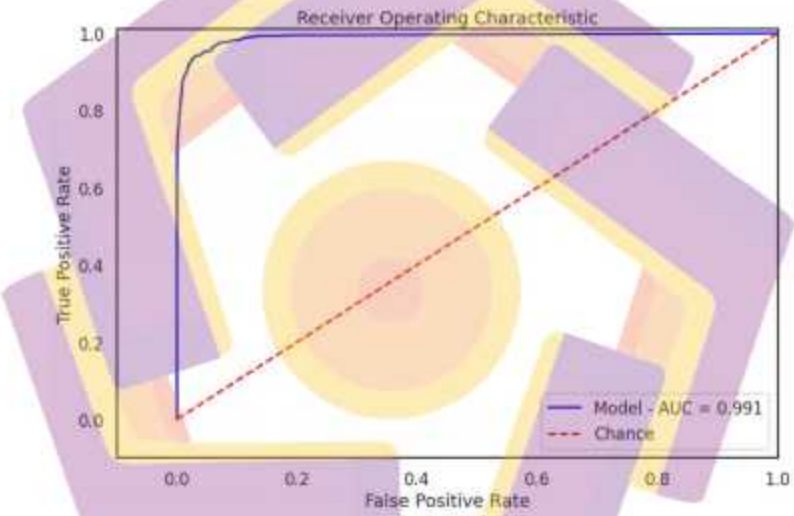
Dalam melakukan klasifikasi menggunakan *Random Forest*, setelah model terbentuk pada data pelatihan maka selanjutnya melakukan pengujian terhadap data validasi untuk melihat akurasi dari model yang didapat. Adapun hasil dari klasifikasi yang dilakukan menggunakan *Random Forest* dapat dilihat pada tabel di bawah ini.

```
Confusion Matrix:
 [[4280   73]
 [  58  550]]
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.98      0.98      4353
           1       0.88      0.90      0.89       608

    accuracy                           0.97      4961
   macro avg       0.93      0.94      0.94      4961
weighted avg       0.97      0.97      0.97      4961

Accuracy Score:  0.9735940334609958
Recall Score: 0.9046052631578947
Precision_score: 0.8828250401284109
f1_score: 0.8935824532900082
AUC:  0.9439176097549418
KS:  0.8878352195098359
```

Gambar 4.10. Hasil Prediksi Data *Testing Random Forest* dengan Data Asli

Pada gambar 4.10. memperlihatkan hasil prediksi dari *data testing Random forest* menggunakan data asli di mana kelas *lumpy* diketahui sebanyak 608 kasus terinfeksi terdapat 550 kasus yang diprediksi secara benar dan 58 kasus tidak tepat prediksinya yang diprediksi oleh model sebagai kasus yang tidak terinfeksi *lumpy* yang didapatkan *class of error* sebesar 9,54%. Sedangkan pada kasus tidak terinfeksi *lumpy*, dari total 4.353 kasus tidak terinfeksi sebanyak 4.280 kasus

terdeteksi secara benar dan sebanyak 73 kasus diprediksi salah sebagai kasus yang terinfeksi dengan *class of error* sebesar 1,68%. Secara keseluruhan tingkat akurasi dari model Random Forest yang sudah terbentuk menggunakan data asli untuk melakukan klasifikasi pada hasil prediksi data validasi sebesar 97,36%. Sedangkan skor Recall yang didapatkan sebesar 90,46% dan skor Presisi-nya sebesar 88,28%. Untuk skor F1 didapatkan nilai sebesar 89,36% dan AUC-nya sebesar 94,39%.



Gambar 4.11. Kurva ROC-AUC Data Asli

Gambar 4.11. menunjukan kurva ROC yang menggambarkan hubungan antara data validasi dan data prediksi. Dari kurva tersebut didapatkan nilai AUC yaitu luas daerah di bawah kurva ROC sebesar 0,991. Hasil klasifikasi berdasarkan nilai AUC tersebut dapat dikatakan baik. Nilai metrik kinerja model yang didapat ini selanjutnya dioptimalkan menggunakan set data yang telah dilakukan teknik *SMOTe*.

4.2.2 Klasifikasi *Random Forest* dengan SMOTe

Skenario pemodelan kedua yaitu melakukan klasifikasi algoritma *Random Forest* dilakukan pada data yang telah dilakukan proses *resampling* dengan SMOTe. Klasifikasi yang dilakukan dengan data validasi berupa data asli sebesar 20% dari total data asli dilakukan dengan cara melakukan pelatihan model menggunakan data yang telah di *resampling* dengan jumlah data yang sama untuk kelas *lumpy* dan *no lumpy*.

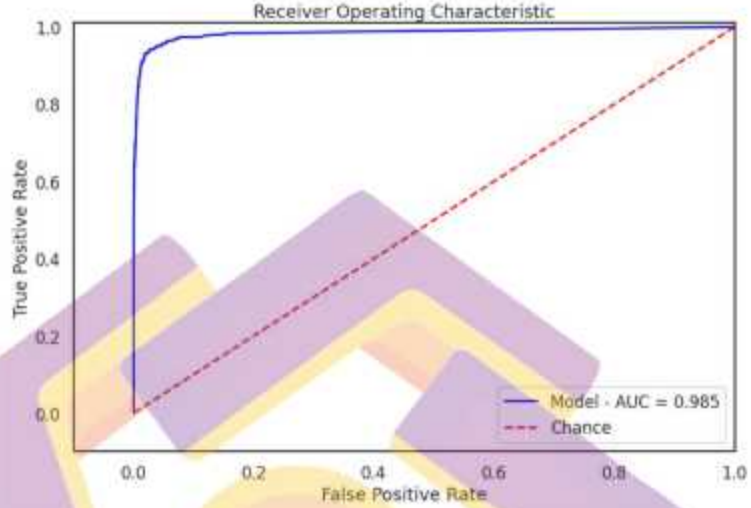
Source code prediksi pada data SMOTe.

```
tpot_classifier.fit(X_res,y_res)
predpotsm = tpot_classifier.predict(X_test)
print("Confusion Matrix: \n", confusion_matrix(y_test, predp
otsm))
print("Classification Report: \n", classification_report(y_t
est, predpotsm))
print('Accuracy score RF: {0:0.4f}'.format(accuracy_score(y_
test, predpotsm)))
print("Recall Score RF:", recall_score(y_test, predpotsm))
print("Precision score RF:", precision_score(y_test, predpot
sm))
print("f1 score RF:", f1_score(y_test, predpotsm))
print("AUC: ", roc_auc_score(y_test, predpotsm))
print("KS: ", ks_2samp(predpotsm[y_test == 0], predpotsm[y_t
est == 1]).statistic)
```

Source code di atas menunjukan proses prediksi yang dilakukan dengan data pelatihan berupa data yang telah memiliki jumlah yang sama untuk masing-masing kelas data dan dilakukan validasi menggunakan data asli. Adapun hasil dari prediksi yang dilakukan dapat dilihat pada gambar di bawah ini.

Gambar 4.12. Hasil Prediksi pada *Data Oversampling SMOTe*

Gambar 4.12. di atas memperlihatkan hasil prediksi dari *data testing Random forest* menggunakan data yang telah di *resampling* dengan teknik SMOTe di mana kelas *lumpy* diketahui sebanyak 607 kasus terinfeksi terdapat 592 kasus yang diprediksi secara benar dan 15 kasus tidak tepat prediksinya yang diprediksi oleh model sebagai kasus yang tidak terinfeksi *lumpy* yang didapatkan *class of error* sebesar 2,47 %. Sedangkan pada kasus tidak terinfeksi *lumpy*, dari total 4.354 kasus tidak terinfeksi sebanyak 4.322 kasus terdeteksi secara benar dan sebanyak 32 kasus diprediksi salah sebagai kasus yang terinfeksi dengan *class of error* sebesar 0,73%. Secara keseluruhan tingkat akurasi dari model *Random Forest* yang sudah terbentuk menggunakan data yang telah di-*resampling* untuk melakukan klasifikasi pada hasil prediksi data validasi sebesar 99%. Sedangkan skor *Recall* yang didapatkan sebesar 98% dan skor Presisi-nya sebesar 95%. Untuk skor F1 didapatkan nilai sebesar 96% dan AUC-nya sebesar 98%.



Gambar 4.13. Kurva ROC-AUC pada Data *SMOTe*

Gambar 4.13. menunjukan kurva ROC yang menggambarkan hubungan antara data validasi dan data prediksi. Dari kurva tersebut didapatkan nilai AUC yaitu luas daerah di bawah kurva ROC sebesar 0,985. Hasil klasifikasi berdasarkan nilai AUC tersebut dapat dikatakan baik.

4.3 Diskusi dan Pembahasan Hasil Penelitian

Bagian ini berisi diskusi dan pembahasan hasil penelitian yang telah dicapai yang meliputi perbandingan terhadap penelitian yang telah dilaksanakan sebelumnya menggunkan dataset yang sama, serta menunjukan kelebihan-kelebihan dan kelemahan-kelemahan yang ada dalam penelitian ini.

4.3.1 Perbandingan dengan Penelitian Terdahulu

Penelitian terdahulu mengenai *lumpy skin disease* telah dilakukan oleh Rai, dkk (2020) yang melakukan pendeteksian LSD berdasarkan gambar pada kulit sapi yang terinfeksi *lumpy* dan *normal skin* menggunakan *Deep Convolutional Neural Network* dengan hasil akurasi sebesar 92,5%. Perbedaan dengan yang penelitian penulis ada pada dataset yang digunakan berbeda jenisnya dimana pada penelitian oleh Rai, dkk menggunakan dataset berbentuk gambar sedangkan penulis menggunakan dataset numerik. Selain itu algoritma pemodelan yang digunakan juga berbeda dimana penelitian terdahulu menggunakan *Deep Convolutional Neural Network* sedangkan penelitian yang dilakukan menggunakan klasifikasi *Random Forest* dengan menerapkan teknik SMOTe

Ghafoor, dkk (2021) dalam penelitiannya tentang pengaplikasian *machine learning* untuk memprediksi risiko mastitis pada sapi dari data sensor AMS yang menggunakan algoritma *random forest* dengan menerapkan *hyperparameter tuning GridSearch* dapat memprediksi risiko mastitis pada sapi dengan akurasi 98,8%. Hal ini tentu saja berbeda dengan penelitian yang dilakukan oleh penulis dimana *hyperparameter tunning* yang digunakan untuk optimalisasi pada penelitian oleh

Ghafoor, dkk adalah *GridSearch, s*edangkan dalam penelitian ini menggunakan *TPOT* untuk optimasinya.

Prediksi otomatis terhadap pola infeksi mastitis pada peternakan sapi perah menggunakan *machine learning* telah dilakukan oleh Hyde, dkk (2020) menggunakan algoritma *Random Forest* yang memiliki hasil *Random Forest* memberikan kinerja model terbaik yang dinilai berdasarkan akurasi, PPV, dan NPV. Perbedaan dengan penelitian yang penulis laksanakan ada pada teknik SMOTe yang dilakukan pada algoritma *Random Forest*.

Workee, dkk (2021) melakukan identifikasi penyakit kulit sapi menggunakan perbandingan tiga teknik filter (median, gaussian, dan gabor filter) dengan hasil akurasi yang dicapai 96,5% di CNN, 93% dengan HOG, dan 98,75% menggunakan fitur *hybrid*. Penelitian yang telah dilakukan dalam bidang *computer vision* untuk pengenalan penyakit kulit, dan klasifikasi sedangkan penelitian yang dilakukan penulis menggunakan data tabular bukan data gambar sehingga algoritma yang digunakan juga berbeda.

Rojas, dkk (2021) dalam penelitian mengenai prediksi risiko penyakit pernapasan sapi pada sapi feedlot 45 hari pertama pasca kedatangan menggunakan algoritma *Logistic Regression, Decision Tree, Random Forest, naïve Bayes* dan *Linear Discriminant* memberikan hasil bahwa AUC pada *Random Forest* memiliki skor tertinggi dibandingkan dengan nilai AUC pada algoritma yang lain sebesar 0,789. Penelitian oleh Rojas, dkk berfokus pada penyakit pernafasan sapi yang menggunakan algoritma *Naïve Bayes, Decision Tree, Random Forest* serta *Logistic Regression*, sedangkan penelitian yang penulis lakukan berfokus pada LSD dengan

menggunakan Algoritma *Random Forest.*

Prediksi penyakit postpartum pada sapi perah menggunakan *machine learning* telah dilakukan oleh Beek, dkk (2018) menunjukan bahwa *Random Forest* merupakan algoritma yang memiliki performa terbaik dibandingkan algoritma lainnya dalam memprediksi penyakit *postpartum* pada sapi perah. Penelitian yang dilakukan untuk memprediksi LSD juga menggunakan algoritma *Random Forest* dengan menambahkan optimasi menggunakan TPOT dan teknik SMOTe.

Pada tahun 2020, Kaler dkk melakukan penelitian yang bertujuan untuk membedakan ketimpangan dalam tiga aktivitas berbeda (berjalan, berdiri, dan berbaring) pada domba menggunakan algoritma *Random Forest*. Algoritma *random forest* bekerja paling baik untuk mengklasifikasi kepincangan dengan akurasi 84,91% dalam berbaring, 81,15% dalam berdiri dan 76,83% dalam berjalan dan secara keseluruhan diklasifikasikan dengan benar lebih dari 80% dalam aktivitas domba. Perbedaan kasus yang digunakan serta optimasi pada algoritma Random Forest menjadi pembeda penelitian ini dengan penelitian yang penulis lakukan.

Penelitian ini menggunakan dataset yang sama dengan penelitian sebelumnya yang telah dilakukan oleh Savafi (2020) yaitu *dataset lumpy skin disease*. Terdapat beberapa perbedaan dalam teknik pemrosesan data maupun dalam pemodelan datanya. Pada pemilihan atribut terdapat perbedaan dari penelitian terdahulu yang dilakukan oleh Savafi (2020), yaitu jumlah atribut yang digunakan dimana Savafi menggunakan 15 atribut dan mengesampingkan atribut-atribut yang berkorelasi sangat tinggi. Dalam penelitian ini, penulis menambahkan

seleksi fitur dengan matrik korelasi sehingga didapatkan 5 atribut yang harus dihilangkan karena memiliki korelasi yang sangat tinggi, bahkan tmn dan tmp, tmp dan tmx berkorelasi 1 yang memiliki arti yang sama. Total atribut yang digunakan dalam penelitian ini menjadi 10 atribut.

Perbedaan selanjutnya yang membedakan penelitian ini dengan penelitian yang dilakukan Safavi yaitu pada penelitian terdahulu tidak dilakukan teknik *resampling* terhadap *dataset* LSD walaupun ada ketidakseimbangan data yang sangat mencolok antara data sapi yang terinfeksi dan yang tidak terinfeksi. Dalam penelitian ini, penulis menerapkan teknik *resampling* menggunakan SMOTe terhadap *dataset* LSD untuk mengatasi permasalahan ketidakseimbangan data yang ada.

Perbedaan lainnya ada pada metode optimasi yang dilakukan pada penelitian ini dengan penelitian sebelumnya (Safavi, 2022) yaitu pada pemilihan metode *hyperparameter tuning*-nya dimana Savafi menggunakan *RandomizedSearchCV* untuk optimasi dengan hasil optimal parameternya *n_estimators = 5000, min_samples_split = 2, bootstrap = True, max_leaf_nodes = 200, class_ weight = {0: 30, 1: 70}, criterion = 'entropy', max_depth = 14.* Sedangkan penelitian ini penulis menggunakan TPOT dengan hasil optimal parameternya *criterion = entropy, max_depth = 500, max_features = auto, min_samples_leaf = 4, min_samples_split = 14, n_estimators = 250.*

Adapun hasil penelitian ini apabila dibandingkan dengan penelitian terdahulu menggunakan dataset yang sama yang telah diteliti oleh Savafi (2020) seperti terlihat pada tabel di bawah ini.

Tabel 4.3. Perbandingan hasil pemodelan penelitian ini dengan penelitian terdahulu

	Penelitian Terdahulu RF+ RandomizedSearchCV	***Penelitian Ini RF+SMOTe***
Accuracy	96%	99%
Precision	89%	95%
Recall	71%	98%
F1-Score	79%	96%
AUC	85%	98%

Pada Tabel 4.3. diatas dapat dilihat bahwa hasil kinerja model Random Forest ini berhasil meningkatkan skor akurasi dari penelitian yang telah dilakukan sebelumnya oleh Safavi (2022) dengan menggunakan *dataset* yang sama pada penggunaan algoritma klasifikasi yang sama, yaitu Random Forest dengan peningkatan sebesar 3% dari 96% menjadi 99%. Skor Presisi mengalami peningkatan 6% dari penelitian sebelumnya 89% menjadi 95%. Skor *Recall* mengalami peningkatan sebanyak 27% dari 71% menjadi 98%. F1-*Score* memiliki peningkatan 17% dari 79% menjadi 96% serta AUC meningkat sebesar 13% dari 85% menjadi 98%.

Dari hasil tersebut didapatkan bahwa seleksi fitur dengan menghilangkan atribut yang memiliki korelasi sangat tinggi sangat berdampak terhadap performa dari kinerja algoritma Random Forest yang dapat dilihat adanya peningkatan pada keseluruhan skor akurasi, *recall*, *f1-score* dan AUC.

Penerapan teknik *resampling* untuk mengatasi ketidakseimbangan data telah dilakukan pada penelitian sebelumnya oleh Mqadi, dkk (2021) mengenai pendekatan *data-point oversampling* berbasis SMOTe untuk memecahkan masalah

ketidakseimbangan data kartu kredit dalam deteksi penipuan keuangan yang berhasil menunjukkan bahwa penggunaan SMOTe terhadap *dataset* yang sangat tidak seimbang tersebut secara signifikan meningkatkan kemampuan untuk memprediksi kelas positif. *Dataset Credit Card* yang berisi 492 data frauds (1) atau 0,172% dari total keseluruhan data sebanyak 284.807 data. Mqadi, dkk (2021) menerapkan *resampling* SMOTe terhadap dataset tersebut sehingga dapat memberikan peningkatan performa model (*Support Vector Machine, Linear Regression, Decission Tree, Random Forest*) yang berupa peningkatan presisi, *Recall*, dan *F1 Score*. Pada pemodelan menggunakan *Random Forest* sendiri terdapat peningkatan *Recall* pada kelas positif dari sebelum SMOTe 53% menjadi 100%, peningkatan Presisi dari 90% menjadi 100% serta peningkatan *F1 Score* dari 67% menjadi 100%. Sedangkan dalam penelitiannya Zhang dan Wang (2011) mengenai metode resampling untuk ketidakseimbangan kelas kartu kredit menyimpulkan bahwa hasil penelitian mengindikasikan bahwa SMOTe merupakan teknik yang lebih efektif dalam penanganan ketidakseimbangan kelas data dibandingan dengan metode *resampling* yang lain.

Mahmudah, dkk (2021) juga menggunakan *oversampling* SMOTe untuk menangani ketidak seimbangan kelas data dalam penelitiannya tentang prediksi penyakit paru obstruktif kronis dari data ekspresi gen. Hasil penelitiannya menunjukan bahwa penggunaan SMOTe dalam menangani ketidakseimbangan kelas data pada algoritma klasifikasi didapatkan bahwa akurasi, specificity, sensitivity dan skor AUC memiliki peningkatan skor daripada sebelumnya.

Pemanfaatan teknik *resampling* dalam penelitian ini menggunakan SMOTe, memiliki dampak bisa meminimalkan *False Negatives* (sapi yang diprediksi sehat tetapi kenyataannya terinfeksi) yang diketahui dari adanya peningkatan metrik *Recall* yang cukup tinggi sebesar 27% dibandingkan dengan penelitian sebelumnya oleh Safavi (2020) yang tidak menerapkan teknik *resampling* terhadap *dataset* LSD.

4.3.2 Perbandingan Evaluasi Kinerja Model

Evaluasi terhadap kinerja model menggunakan *Confusion Matrix* untuk mengetahui akurasi, presisi, *recall, F1-Score* dari model. Berdasarkan skenario pengujian penggunaan klasifikasi Random Forest terhadap *dataset* LSD terhadap data asli serta melakukan klasifikasi Random Forest dengan teknik *resampling* SMOTe. Hasil dari pengujian kedua eksperimen tersebut seperti pada tabel berikut ini.

Tabel 4.4. Perbandingan hasil pemodelan Random Forest Sebelum dan Sesudah Menerapkan *SMOTe*

	Random Forest	***Random Forest+SMOTe***
Accuracy	97%	99%
Recall	90%	98%
Precision	88%	95%
F1-Score	89%	96%
AUC	94%	98%

Hasil dari evaluasi model pada Tabel 4.4 menggunakan *Confusion Matrix* dengan dua skenario pengujian mendapatkan hasil bahwa terjadi peningkatan

sebesar 2 % pada akurasi dari 97% menjadi 99%. Peningkatan F1-Score sebanyak 7% dari 89% menjadi 96%. Presisi mengalami peningkatan sebesar 7% dari 88% menjadi 95%. Skor AUC mengalami peningkatan skor sebanyak 4% dari sebelumnya 94% menjadi 98%. Kombinasi antara Random Forest dan SMOTe telah berhasil mengurangi *False Negatives*, tanpa harus mengorbankan terlalu banyak *False Positivex* (tetapi tetap harus menguji sapi sehat) dengan peningkatan skor recall sebesar 8% dari 90 % menjadi 98%. Akurasi 99% menunjukan bahwa model dapat mengklasifikasikan sapi yang tidak terinfeksi *lumpy* dengan sangat baik.

Penelitian ini telah dipublikasikan ke dalam seminar international pada *5th International Conference on Information and Communications Technology (ICOIACT)* 2022 dengan judul yang sama yaitu *"Lumpy Skin Disease Prediction Based on Meteorological and Geospatial Features using Random Forest Algorithm with Hyperparameter Tuning"* serta publikasi pada jurnal Resti terindeks Sinta-2 yang telah terbit pada Vol 6 No 4 (2022) bulan Agustus 2022 dengan judul *"Applying Different Resampling Strategies In Random Forest Algorithm To Predict Lumpy Skin Disease"*.

BAB V

PENUTUP

5.1. Kesimpulan

a) Pengimplementasian SMOTe terhadap *dataset* LSD yang dilakukan dengan cara mensintesis kelas data minoritas sejumlah kelas data mayoritas telah berhasil menyeimbangkan jumlah data yang digunakan dalam pelatihan sehingga hasil kinerja model tidak menjadi bias.

b) Metode *feature reduction* dan SMOTe yang digunakan mampu meningkatkan akurasi prediksi kelas minoritas yang dapat mengoptimasi kinerja *Random Forest Classifier* tanpa menyebabkan adanya informasi yang hilang, menghindari terjadinya overfitting, membangun wilayah keputusan yang lebih besar, serta terbukti dengan adanya peningkatan skor kinerja model jika dibandingkan dengan penelitian Savafi (2022).

c) Model yang dirancang telah memberikan kinerja yang baik *(finiteness)* dalam merespon *dataset* yang dievaluasi dengan akurasi sebesar 99% dimana model dapat dengan baik mengklasifikan sapi yang terinfeksi maupun yang tidak terinfeksi.

d) Kombinasi antara Random Forest dan SMOTe telah berhasil mengurangi *False Negatives*, tanpa harus mengorbankan terlalu banyak *False Positives* (tetapi tetap harus menguji sapi sehat) terbukti dengan adanya peningkatan nilai *recall* sebesar 27% dari 71% menjadi 98% dibandingkan penelitian Savafi (2022) yang tidak mengimplementasikan SMOTe.

5.2. Saran

a) Peningkatan *Recall* atau minimalisasi *False Negative* masih dapat dilakukan dengan menggunakan teknik *resampling* yang lainnya seperti Random Over Sampling, BorderLine Smote, KMeans Smote, SVM Smote, ADASYN ataupun Smote-NC.

b) Pemilihan parameter awal pada klasifikasi Random Forest dapat sesuaikan kembali untuk mendapatkan metrik yang lebih baik lagi.

c) Pemilihan model *machine learning* dimungkinkan untuk menggunakan metode klasifikasi yang baru maupun metode *deep learning*.

DAFTAR PUSTAKA

PUSTAKA BUKU

Breiman, L. *et al.* (2017) *Classification And Regression Trees*. New York: Routledge. doi: 10.1201/9781315139470.

Breiman, L. E. O. (2001) 'Random Forest', in *Machine Learning*, pp. 5–32.

Géron, A. (2019) *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 2nd edn. Seoul, Korea: Hanbit Media, Inc.

Hutter, F. and Kotthoff, L. (2019) *Automated Machine Learning*. Edited by J. Vanschoren. Springer.

Ibrahim, J., Chen, M.-H., & Sinha, D. (2009). The elements of statistical learning: data mining, inference, and prediction. In *The Elements of Statistical Learning* (Vol. 27, Issue 2). springer. http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf

Safavi, E. A. (2022). Assessing machine learning techniques in forecasting lumpy skin disease occurrence based on meteorological and geospatial features. *Tropical Animal Health and Production*, 1–11. https://doi.org/10.1007/s11250-022-03073-2

Ibrahim, J., Chen, M.-H., & Sinha, D. (2009). The elements of statistical learning: data mining, inference, and prediction. In *The Elements of Statistical Learning* (Vol. 27, Issue 2). springer. http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf

Ramsay, J. O., Berge, J. Ten, & Styan, G. P. H. (1984). Matrix correlation. *Psychometrika*, *49*(3), 403-423.

Safavi, E. A. (2022). Assessing machine learning techniques in forecasting lumpy skin disease occurrence based on meteorological and geospatial features. *Tropical Animal Health and Production*, 1–11.

https://doi.org/10.1007/s11250-022-03073-2

Ibrahim, J., Chen, M.-H., & Sinha, D. (2009). The elements of statistical learning: data mining, inference, and prediction. In *The Elements of Statistical Learning* (Vol. 27, Issue 2). springer. http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf

Kahl, C., & Günther, M. (2008). Complete the correlation matrix. *From Nano to Space*, 239–244. https://doi.org/10.1007/978-3-540-74238-8_17

Ramsay, J. O., Berge, J. Ten, & Styan, G. P. H. (1984). Matrix correlation. *Psychometrika*, *49*(3), 403–423.

Safavi, E. A. (2022). Assessing machine learning techniques in forecasting lumpy skin disease occurrence based on meteorological and geospatial features. *Tropical Animal Health and Production*, 1–11. https://doi.org/10.1007/s11250-022-03073-2

Ibrahim, J., Chen, M.-H., & Sinha, D. (2009). The elements of statistical learning: data mining, inference, and prediction. In *The Elements of Statistical Learning* (Vol. 27, Issue 2). springer. http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf

Kahl, C., & Günther, M. (2008). Complete the correlation matrix. *From Nano to Space*, 239–244. https://doi.org/10.1007/978-3-540-74238-8_17

Ramsay, J. O., Berge, J. Ten, & Styan, G. P. H. (1984). Matrix correlation. *Psychometrika*, *49*(3), 403–423.

Safavi, E. A. (2022). Assessing machine learning techniques in forecasting lumpy skin disease occurrence based on meteorological and geospatial features. *Tropical Animal Health and Production*, 1–11. https://doi.org/10.1007/s11250-022-03073-2

Gorunescu, F. (2010). *Data Mining Concepts, Models and Techiniques*. Springer International Publishing. https://doi.org/https://doi.org/10.1007/978-3-642-19721-5

Ibrahim, J., Chen, M.-H., & Sinha, D. (2009). The elements of statistical learning: data mining, inference, and prediction. In *The Elements of Statistical Learning* (Vol. 27, Issue 2). springer. http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf

Kahl, C., & Günther, M. (2008). Complete the correlation matrix. *From Nano to Space*, 239–244. https://doi.org/10.1007/978-3-540-74238-8_17

Ramsay, J. O., Berge, J. Ten, & Styan, G. P. H. (1984). Matrix correlation. *Psychometrika*, *49*(3), 403–423.

Safavi, E. A. (2022). Assessing machine learning techniques in forecasting lumpy skin disease occurrence based on meteorological and geospatial features. *Tropical Animal Health and Production*, 1–11. https://doi.org/10.1007/s11250-022-03073-2

Gorunescu, F. (2010). *Data Mining Concepts, Models and Techiniques*. Springer International Publishing. https://doi.org/https://doi.org/10.1007/978-3-642-19721-5

Ibrahim, J., Chen, M.-H., & Sinha, D. (2009). The elements of statistical learning: data mining, inference, and prediction. In *The Elements of Statistical Learning* (Vol. 27, Issue 2). springer. http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf

Kahl, C., & Günther, M. (2008). Complete the correlation matrix. *From Nano to Space*, 239–244. https://doi.org/10.1007/978-3-540-74238-8_17

Ramsay, J. O., Berge, J. Ten, & Styan, G. P. H. (1984). Matrix correlation. *Psychometrika*, *49*(3), 403–423.

Safavi, E. A. (2022). Assessing machine learning techniques in forecasting lumpy skin disease occurrence based on meteorological and geospatial features. *Tropical Animal Health and Production*, 1–11. https://doi.org/10.1007/s11250-022-03073-2

Wang, W., & Zhang, L. (2011). A Re-sampling Method for Class Imbalance Learning with Credit Data. *International Conference of Information Technology, Computer Engineering and Management Sciences*, 393–397. https://doi.org/10.1109/ICM.2011.34

Gorunescu, F. (2010). *Data Mining Concepts, Models and Techiniques*. Springer International Publishing. https://doi.org/https://doi.org/10.1007/978-3-642-19721-5

Ibrahim, J., Chen, M.-H., & Sinha, D. (2009). The elements of statistical learning: data mining, inference, and prediction. In *The Elements of Statistical Learning* (Vol. 27, Issue 2). springer.

http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf

Kahl, C., & Günther, M. (2008). Complete the correlation matrix. *From Nano to Space*, 239–244. https://doi.org/10.1007/978-3-540-74238-8_17

Ramsay, J. O., Berge, J. Ten, & Styan, G. P. H. (1984). Matrix correlation. *Psychometrika*, *49*(3), 403-423.

Safavi, E. A. (2022). Assessing machine learning techniques in forecasting lumpy skin disease occurrence based on meteorological and geospatial features. *Tropical Animal Health and Production*, 1–11. https://doi.org/10.1007/s11250-022-03073-2

Wu, C., Tzeng, G., Goo, Y.-J., & Fang, W.-C. (2007). A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcy. *Expert Systems with Applications 32*, *32*, 397–408. https://doi.org/10.1016/j.eswa.2005.12.008

Kuhn, M. and Johnson, K. (2013) *Applied Predictive Modeling*. 26th edn. Springer.

Naqa, I. El and Murphy, M. J. (2015) 'What Is Machine Learning ?', in *Machine learning in radiation oncology*. Springer, pp. 3–11. doi: 10.1007/978-3-319-18305-3.

Tuppurainen, E. S. M., Babiuk, S. and Klement, E. (2018) *Lumpy Skin Disease*. Springer International Publishing. doi: 10.1007/978-3-319-92411-3_4.

Zhou, Z. (2016) *Machine Learning*. Beijing, China: Tsinghua University Press.

PUSTAKA MAJALAH, JURNAL ILMIAH ATAU PROSIDING

Agarwal, A., Sharma, P., Alshehri, M., Mohamed, A. A., & Alfarraj, O. (2021). *Classification model for accuracy and intrusion detection using machine learning approach*. 1–22. https://doi.org/10.7717/peerj-cs.437

Alemayehu, G., Zewde, G., & Admassu, B. (2013). Risk Assessments of Lumpy Skin Diseases in Borena Bull Market Chain and Its Implication for Livelihoods and International Trade. *Trop Anim Health Prod*, 1153–1159. https://doi.org/10.1007/s11250-012-0340-9

Aljameel, S. S., Khan, I. U., Aslam, N., Aljabri, M., & Alsulmi, E. S. (2021). Machine Learning-Based Model to Predict the Disease Severity and Outcome in COVID-19 Patients. *Hindawi Scientific Programming*. https://doi.org/10.1155/2021/5587188

Alkhamis, M. A. K. V. (2016). *Spatial and Temporal Epidemiology of Lumpy Skin Disease in the Middle East, 2012–2015*. *3*(March), 1–12. https://doi.org/10.3389/fvets.2016.00019

Allepuz, A., Casal, J., & Beltrán-Alcrudo, D. (2019). Spatial Analysis of Lumpy Skin Disease in Eurasia-Predicting Areas at Risk for Further Spread within The Region. *Transboundary and Emerging Diseases*, 0–3. https://doi.org/10.1111/tbed.13090

Anggara, D. S., & Abdillah, C. (2019). *Modul Metode Penelitian*. Universitas Pamulang.

Beek, S. Van Der, Layer, G., & Gmbh, G. (2018). Prediction of postpartum diseases of dairy cattle using machine learning. *Proceedings of the World Congress on Genetics Applied to Livestock Production*, *11.104*.

Bellinger, C., Drummond, C., & Japkowicz, N. (2017). Manifold-based synthetic oversampling with manifold conformance estimation. *Mach. Learn.*, *107*, 605–637.

Bramer, M., Stahl, F., & Gaber, M. M. (2013). Scaling up Data Mining Techniques to Large Datasets Using Parallel and Distributed Processing. In *Business Intelligence and Performance Management* (pp. 243–259). Springer. https://doi.org/10.1007/978-1-4471-4866-1

Carrizosa, E., Molero, C., Dolores, R., & Morales, R. (2021). Mathematical optimization in classification and regression trees. *TOP*, *29*(1), 5–33. https://doi.org/10.1007/s11750-021-00594-1

Chawla, N., K, B., L, H., & W, K. (2002). SMOTE: Synthetic Minority Over–Sampling Technique. *Journal of Artificial Intelligence Re-Search*, *16*(1), 321–357.

Chihota, C. M. (2003). Attempted mechanical transmission of lumpy skin disease virus by biting insects. *Medical and Veterinary Entomology*, *17*, 294–300.

Dimitriadis, N. (2020). Applying Topic Modelling Algorithms on Twitter messages in Greek language. *Ikee.Lib.Auth.Gr*. http://ikee.lib.auth.gr/record/324006/files/Dimitriadis-2158.pdf

Elgeldawi, E., Sayed, A., Galal, A. R., & Zaki, A. M. (2021). Hyperparameter Tuning for Machine Learning Algorithms Used for Arabic Sentiment Analysis. *Informatics*, *8*(79), 1–21.

Garcia, R. (2017). "One Wefare": a framework to support the implementation of OIE animal wefare standards. *OIE*, 3–13. https://doi.org/10.20506/bull.2017.1.2589

García, S., Luengo, J., & Herrera, F. (2015). Feature selection : A Data Perspective. *Intelligent Systems Reference Library*, *72*(6), 163–193. https://doi.org/10.1007/978-3-319-10247-4_7

Ghafoor, N. A. (2021). MasPA : A Machine Learning Application to Predict Risk of Mastitis in Cattle from AMS Sensor Data. *AgriEngineering*, *3*, 575–583.

Hong, C. S., & Gyu, T. (2021). TPR-TNR plot for confusion matrix. *Communications for Statistical Applications and Methods*, *28*(2), 161–169. https://doi.org/https://doi.org/10.29220/CSAM.2021.28.2.161

Huang, X., Chen, S., Chen, H., Hu, L., Wen, L., Wei, F., & Chen, K. (2021). ROC Curve Analysis of the Sensitivity and Specificity of Biochemical Detection of Intrahepatic Cholestasis during Pregnancy. *Zeitschrift Fur Geburtshilfe Und Neonatologie*, *225*(4), 327–332. https://doi.org/10.1055/a-1299-2298

Hyde, R. M., Down, P. M., Bradley, A. J., Breen, J. E., Hudson, C., Leach, K. A., & Green, M. J. (2020). Automated prediction of mastitis infection patterns in dairy herds using machine learning. *Scientific Reports*, *10*(4289), 1–8. https://doi.org/10.1038/s41598-020-61126-8

Ince, O. B. (2020). Analyzing risk factors for lumpy skin disease by a geographic information system (GIS) in Turkey. *Journal of the Hellenic Veterinary Medical Society*, *70*(4), 1797–1804. https://doi.org/10.12681/jhvms.22222

Iskandar, A. F., Utami, E., & Prasetio, A. B. (2020). Word Analysis of Indonesian Keirsey Temperament. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *14*(4), 365. https://doi.org/10.22146/ijccs.58595

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Big Data*, *6*, 1–54.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245).

Kaler, J., Mitsch, J., Vázquez-diosdado, J. A., Bollard, N., Dottorini, T., & Ellis, K. A. (2020). Automated detection of lameness in sheep using machine learning approaches : novel insights into behavioural differences among lame and non-lame sheep. *Royal Society Open Science*, *7*(190824). https://doi.org/http://dx.doi.org/10.1098/rsos.190824

Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making*, *11*(51). https://doi.org/10.1186/1472-6947-11-51

Liang, R., Lu, Y., Qu, X., Su, Q., Li, C., Xia, S., Liu, Y., Zhang, Q., Cao, X., Chen, Q., & Niu, B. (2020). Prediction for global African swine fever outbreaks based on a combination of random forest algorithms and meteorological data. *Transboundary and Emerging Diseases*, *67*(September 2019), 935–946. https://doi.org/10.1111/tbed.13424

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(December), 18–22.

Lojkic, I. (2018). Complete Genome Sequence of a Lumpy Skin Disease Virus Strain Isolated from the Skin of a Vaccinated Animal. *American Society for Microbiology*, 1–2. https://doi.org/https://doi.org/10.1128/genomeA .00482-18

Machado, G., Korennoy, F., Alvarez, J., Risso, C. P., Perez, A., Vanderwaal, K., & Carolina, N. (2019). Mapping changes in the spatiotemporal distribution of lumpy skin disease virus. *Transbound Emerg Disease*, 1–13. https://doi.org/10.1111/tbed.13253

Mammadova, N., & Keskin, E. (2013). *Application of the Support Vector Machine to Predict Subclinical Mastitis in Dairy Cattle*. *2013*.

Miao, J., & Zhu, W. (2021). Precision-recall curve (PRC) classification trees. *Evolutionary Intelligence*. https://doi.org/10.1007/s12065-021-00565-2

Michalewicz, Z. (1992). *Genetic algorithms + data structures = evolution programs*. Springer-Verlag.

Molla, W., Frankena, K., Gari, G., Kidane, M., Shegu, D., & Jong, M. C. M. de. (2018). Seroprevalence and risk factors of lumpy skin disease in Ethiopia. *Preventive Veterinary Medicine*. https://doi.org/10.1016/j.prevetmed.2018.09.029

Molla, W., Jong, M. C. M. de, & Frankena, K. (2017). Temporal and spatial distribution of lumpy skin disease outbreaks in Ethiopia in the period 2000 to 2015. *BMC Veterinary Research*, 1–9. https://doi.org/10.1186/s12917-017-1247-5

Morris et al. (1930). *Pseudo-urticaria*.

Namazi, F., & Tafti, A. K. (2021). Lumpy skin disease , an emerging transboundary viral disease : A review. *Vet Med Sci*, *7*, 888–896. https://doi.org/10.1002/vms3.434

Ochwo, S., Vanderwaal, K., Munsey, A., Nkamwesiga, J., Ndekezi, C., Auma, E., & Mwiine, F. N. (2019). Seroprevalence and risk factors for lumpy skin disease virus seropositivity in cattle in Uganda. *BMC Veterinary Research*, *15*(236), 1–9.

Orynbassar, A., Sapazhanov, Y., Kadyrov, S., & Lyublinskaya, I. (2022). Applications of ROC Curves Analysis for Predicting Students' Success in a Course based on Prerequisite Grades. *Mathematics*, *10*(12), 1–11. https://doi.org/10.3390/math10122084

Pakpahan, M., Amruddin, A., Sihombing, R. M., Siagian, V., Kuswandi, S., Arifin, R., Mukhoirotin, M., Karwanto, K., Tasrim, I. W., Kato, I., Subakti, H., & Aswan, N. (2022). *Metodologi Penelitian* (A. Karim, Ed.). Yayasan Kita Menulis.

Pande, A. (2019). *An Efficient Approach to Fruit Classification and Grading using Deep Convolutional Neural Network*. 1–7.

Pérez, S., Pablo, F., Camblor, M., & Filzmoser, P. (2020). Visualizing the decision rules behind the ROC curves : understanding the classification process. *AStA Advances in Statistical Analysis*, *0123456789*. https://doi.org/10.1007/s10182-020-00385-2

Rahayu, S., Jaya Purnama, J., Baroqah Pohan, A., Septia Nugraha, F., Nurdiani, S., & Hadianti, S. (2020). Prediction of Survival of Heart Failure Patients Using Random Forest. *Jurnal Pilar Nusa Mandiri*, *16*(2), 255–260.

Rai, G., Hussain, A., Kumar, A., Rai, G., Hussain, A., Kumar, A., & Nijhawan, R. (2020). A Deep Learning Approach to Detect Lumpy Skin Disease in Cows. *EasyChair Preprint*.

Ramsay, J. O., Berge, J. ten, & Styan, G. P. H. (1984). Matrix correlation. *Psychometrika*, *49*(3), 403–423.

Rodriguez, J. D., Perez, A., & Lozano, J. A. (2010). Sensitivity Analysis of k -Fold Cross Validation in Prediction Error Estimation. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, *32*(3), 569–575. https://doi.org/10.1109/TPAMI.2009.187

Rojas, H. A., White, B. J., Amrine, D. E., & Larson, R. L. (2022). Predicting Bovine Respiratory Disease Risk in Feedlot Cattle in the First 45 Days Post Arrival. *MDPI*.

Safavi, E. A. (2022). Assessing machine learning techniques in forecasting lumpy skin disease occurrence based on meteorological and geospatial features. *Tropical Animal Health and Production*, 1–11. https://doi.org/10.1007/s11250-022-03073-2

Samuel, A. L. (2000). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *44*(1).

Sendow, I., Assadah, N. S., Ratnawati, A., Dharmayanti, N., & Saepulloh, M. (2021). Lumpy Skin Disease : Ancaman Penyakit Emerging bagi Status Kesehatan Hewan Nasional (Lumpy Skin Disease : Emerging Diseases Threats for National Animal Health Status). *WARTAZOA*, *31*(2), 85–96.

Shelke, M. S., Deshmukh, P. R., & Shandilya, P. V. K. (2017). A Review on Imbalanced Data Handling Using Undersampling and Oversampling Technique. *International Journal of Recent Trends in Engineering & Research Research (IJRTER)*, *3*(4), 444–449.

Sprygin, A., Pestova, Y., Wallace, D. B., Tuppurainen, E., & Kononov, A. v. (2019). Transmission of lumpy skin disease virus : A short review. *Virus Research*, *269*(May), 197637. https://doi.org/10.1016/j.virusres.2019.05.015

Stern, R. H. (2021). Interpretation of the Area Under the ROC Curve for Risk Prediction Models. *ArXiv Preprint*, *2102.11053*.

Taser, P. Y. (2021). Application of Bagging and Boosting Approaches Using Decision Tree-Based Algorithms in Diabetes Risk Prediction †. *MDPI*, *74*(6).

Thejas, G. S., Hariprasad, Y., Iyengar, S. S., Sunitha, N. R., & Badrinath, P. (2022). Machine Learning with Applications An extension of Synthetic Minority Oversampling Technique based on Kalman filter for imbalanced datasets. *Machine Learning with Applications*, *8*(July 2021), 100267. https://doi.org/10.1016/j.mlwa.2022.100267

Tuppurainen, E. S. M., Babiuk, S., & Klement, E. (2018). *Lumpy Skin Disease*. Springer International Publishing. https://doi.org/10.1007/978-3-319-92411-3_4

Utami, E., Oyong, I., Raharjo, S., Dwi Hartanto, A., & Adi, S. (2021). Supervised learning and resampling techniques on DISC personality classification using Twitter information in Bahasa Indonesia. *Applied Computing and Informatics*. https://doi.org/10.1108/ACI-03-2021-0054

Wang, S., & Yao, X. (2013). Using Class Imbalance Learning for Software Defect Prediction. *IEEE TRANSACTIONS ON RELIABILITY*, *62*(2), 434–443.

Wang, W., & Zhang, L. (2011). A Re-sampling Method for Class Imbalance Learning with Credit Data. *International Conference of Information Technology, Computer Engineering and Management Sciences*, 393–397. https://doi.org/10.1109/ICM.2011.34

Widyastuti, N., & Hamzah, A. (2007). PENGGUNAAN ALGORITMA GENETIKA DALAM PENINGKATAN KINERJA FUZZY CLUSTERING UNTUK (Application of Genetic Algorithm to Enhance the Performance of Clustering. *Berkala MIPA*, *17*(2), 1–14.

Workee, G. (2021). *Cattle skin diseases identification model using machine learning approach*. Bahir Dar University.

Wu, C., Tzeng, G., Goo, Y.-J., & Fang, W.-C. (2007). A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcy. *Expert Systems with Applications 32*, *32*, 397–408. https://doi.org/10.1016/j.eswa.2005.12.008

Xu, J., Zhang, Y., & Miao, D. (2020). Three-way confusion matrix for classification : A measure driven view. *Information Sciences*, *507*, 772–794. https://doi.org/10.1016/j.ins.2019.06.064

PUSTAKA ELEKTRONIK

Epizooties, [OIE] Office International des (2017) *Manual of diagnostic tests and vaccines for terrestrial animals, chapter 2.4.14, Lumpy skin disease*. Available

at: http://web.oie.int/eng/normes/MMANUAL/A_Index.htm (Accessed: 28 March 2022).

Pertanian, K. (2022) *Kementan Siapkan Sumberdaya Tangani Lumpy Skin Disease Pada Sapi Di Riau.* Available at: http://ditjenpkh.pertanian.go.id/kementan-siapkan-sumberdaya-tangani-lumpy-skin-disease-pada-sapi-di-riau (Accessed: 7 April 2022).

LAMPIRAN