JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 8, No. 2, Juni 2023, Pp. 448-458

# KLASIFIKASI LAJU PERNAFASAN DAN SATURASI OKSIGEN MENGGUNAKAN METODE REGRESI LOGISTIK

**Alfi Zahra Hafizhah*[1)], Sinung Suakanto[2)], Riska Yanu Fa'rifah[3)], Edi Triono Nuryatno[4)]**

1. Universitas Telkom, Indonesia
2. Universitas Telkom, Indonesia
3. Universitas Telkom, Indonesia
4. The University of Western Australia, Australia

**Article Info**

**Kata Kunci:** saturasi oksigen, laju pernapasan, regresi logistik, klasifikasi kondisi pernapasan, downsampling, undersampling

**Keywords:** oxygen saturation, respiratory rate, logistic regression, classification of respiratory condition, downsampling, undersampling

**Article history:**
Received 10 January 2023
Revised 17 January 2023
Accepted 12 February 2023
Available online 1 June 2023

**DOI :**
https://doi.org/10.29100/jipi.v8i2.3481

* Corresponding author.
Corresponding Author
E-mail address:
alfizahrahafizhah7@gmail.com

**ABSTRAK**

Saturasi oksigen dan laju pernapasan adalah dua parameter dasar yang digunakan untuk menilai kondisi pasien, khususnya pernapasan. Gagal jantung dan COVID-19 adalah beberapa penyakit yang berhubungan dengan dua parameter ini. Gagal jantung memiliki gejala pernapasan spesifik seperti nyeri dada dan sesak napas, yang disebabkan oleh ketidaknormalan pada saturasi oksigen dan laju pernapasan. COVID-19 merupakan penyakit yang baru ditemukan pada tahun 2019 dan penyakit ini juga memiliki keterkaitan yang dekat dengan pernapasan. Jika terinfeksi, COVID-19 dapat menyebabkan acute respiratory distress syndrome (ARDS), pneumonia, dan permasalahan dengan organ tubuh lainnya, yang dapat menyebabkan kematian bagi penderitanya. Maka dari itu, kedua parameter ini sangat penting untuk menentukan kondisi pernapasan pasien. Penelitian ini bertujuan untuk membangun sebuah model regresi logistik untuk mengklasifikasikan kondisi pernapasan pasien menggunakan saturasi oksigen dan laju pernapasan sebagai parameter. Regresi logistik digunakan karena kecocokan dari kelebihan model dengan data yang digunakan dalam penelitian dan algoritma ini dapat menjelaskan pengaruh parameter-parameter independen yang digunakan terhadap parameter dependennya. Kemudian model ini akan di evaluasi menggunakan metode F1-Macro. Penyelesaian penelitian menggunakan CRISP-DM metodologi, serta mempersiapkan data menggunakan metode downsampling dan mengategorikan nilai dari variabel-variabel untuk mendapatkan hasil model yang lebih baik. Akurasi dari model testing adalah 87.5%, sementara akurasi evaluasi menggunakan F1-Macro adalah 87%. Hasil dari penelitian ini juga sudah sesuai dengan teori medis yang dilihat dari interpretasi koefisien saturasi oksigen dan laju pernapasan.

**ABSTRACT**

Oxygen saturation and respiratory rate are two basic parameters used to assess the patient's condition, especially respiration. Heart failure and COVID-19 are some of the diseases related to these two parameters. Heart failure has specific respiratory symptoms such as sudden chest pains and shortness of breath, which are caused by an abnormality in oxygen saturation and respiratory rate. COVID-19 is a new disease found in 2019, and this disease also has a close relationship with respiratory. If infected, COVID-19 can cause acute respiratory distress syndrome (ARDS), pneumonia, and problems with other body organs, which can cause death for the sufferer. Therefore, these two parameters are very important to determine the patient's respiratory condition. This study aims to build a logistic regression model for classifying the patient's respiratory condition using oxygen saturation and respiratory rate as parameters. Logistic regression is used because of the suitability of the model's advantages with the data and this algorithm can explain the effect of the independent parameters used on the dependent parameter. Then this model will be evaluated using the F1-Macro method. This study uses the CRISP-DM methodology and prepares data using the downsampling methodology and categorizing values of the variables to get a better model result. The accuracy of the testing model is 87.5%, while the evaluation accuracy using F1-Macro is 87%. This study's results are also

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 8, No. 2, Juni 2023, Pp. 448-458

already appropriate to existing medical theories regarding oxygen saturation and respiratory rate coefficient values.

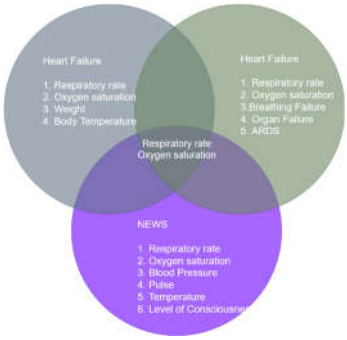
## I. PENDAHULUAN

Saturasi oksigen dan laju pernapasan adalah dua parameter medis yang umum digunakan untuk melihat kondisi pernapasan pasien. Saturasi oksigen merupakan persentase hemoglobin yang terikat oksigen di dalam darah. Saturasi oksigen normal berada dalam rentang 95%-100%. Dampak pada tubuh yang kekurangan saturasi oksigen atau $SpO_2$ adalah nafas yang lebih pendek (sesak nafas/*dyspnea*) sebagai respons paru-paru untuk meningkatkan oksigen dalam darah, sehingga pada sistem pernapasan dapat menyebabkan laju pernapasan atau yang biasa disebut *respiratory rate* atau laju pernapasan menjadi tidak teratur karena kekurangan oksigen di dalam darah [1]. *Respiratory rate* adalah laju pernapasan dalam satu menit atau 60 detik. Hasil pengukuran *respiratory rat*e dapat dipengaruhi oleh banyak faktor [1].

Salah satu penyakit yang berkaitan erat dengan saturasi oksigen dan laju pernapasan adalah *Heart Failure* atau gagal jantung. Gagal jantung merupakan penyakit tidak menular yang mengakibatkan kematian tertinggi di dunia dan Indonesia menempati peringkat kematian tertinggi akibat gagal jantung di Asia dengan jumlah 371 ribu jiwa penderita. Gejala utama dari gagal jantung adalah nyeri dada yang timbul secara mendadak dan sesak nafas. Sekitar 75% – 89% pasien gagal jantung menunjukkan adanya penurunan saturasi oksigen [2].

Penyakit lainnya yang juga berkaitan dengan kedua parameter yang digunakan dalam penelitian ini adalah COVID-19. *Severe Acute Respiratory Syndrome Coronavirus* 2 (SARS-Cov-2) atau yang biasa dikenal sebagai COVID-19 merupakan jenis virus *corona* yang baru ditemukan di Wuhan, China pada Desember 2019. Apabila terinfeksi, COVID-19 dapat menimbulkan berbagai komplikasi penyakit terutama gangguan pada saluran pernapasan akut, *acute respiratory distress syndrome* (ARDS), *pneumonia*, dan juga permasalahan pada organ lainnya yang dapat mengakibatkan kematian bagi penderitanya. Pasien positif di Indonesia per 13 Oktober 2021 adalah sebanyak 4,231,046 dan secara global 225 negara per 13 Oktober 2021 sebanyak 238,521,855 [3]. Pasien COVID-19 yang datang ke rumah sakit memiliki gejala gangguan pernapasan dengan kondisi pengukuran saturasi oksigen <93%, laju pernapasan >30x/menit, ARDS, dan kegagalan pada organ [4].

Selain sebagai parameter medis untuk penyakit tertentu, saturasi oksigen dan laju pernapasan juga merupakan dua dari total 6 parameter yang digunakan untuk *National Early Warning Score* (NEWS). NEWS merupakan sistem parameter digunakan oleh *The National Health Service* untuk menilai perburukkan kondisi pasien dan untuk memprediksi kematian pasien rawat inap atau ICU [5]. Dapat terlihat bahwa selain berkaitan erat dengan penyakit yang berhubungan dengan pernapasan, kedua parameter medis ini juga digunakan sebagai parameter acuan dalam menilai kondisi pasien.




Gambar 1 Diagram Venn Gejalan Penyakit Pernapasan

Dari penjelasan di atas serta gambar 1, dapat dilihat bahwa saturasi oksigen dan laju pernapasan merupakan dua parameter medis yang berhubungan dengan penyakit pernapasan; beberapa di antaranya adalah COVID-19 dan gagal jantung, serta juga digunakan sebagai parameter untuk menilai kondisi pasien. Oleh karena itu penulis ingin membangun model regresi logistik untuk mengklasifikasikan kondisi pernapasan pasien menggunakan dua parameter pengukuran medis yang berkaitan erat dengan pernapasan, yaitu saturasi oksigen dan laju pernapasan. Kondisi pernapasan akan diklasifikasikan dalam dua kelompok, yaitu kondisi pernapasan aman dan tidak aman. *Dataset* yang digunakan dalam penelitian ini adalah *dataset* jurnal penelitian mengenai model prediksi kematian pasien rawat inap dengan gagal jantung, yang di ambil dari Kaggle.com.

Regresi logistik merupakan salah satu model dari *supervised learning* untuk memprediksi variabel dependen yang dapat dikategorikan, dan biasanya memiliki variabel independen yang berbentuk skala numerik *continuous* atau kategori [6]. Model paling popular untuk menangani data biner adalah regresi logistik [7]. Selain itu, model ini juga

dapat membantu menjelaskan besarnya pengaruh variabel independen yang digunakan terhadap variabel dependen. Sehingga dengan menggunakan algoritma ini, penelitian ini dapat menggambarkan dengan jelas besarnya pengaruh kedua parameter, saturasi oksigen dan laju pernapasan, terhadap kondisi pernapasan pasien. Pada tabel I terdapat beberapa penelitian terdahulu yang menggunakan dan membahas mengenai regresi logisitik dan/atau laju pernapasan dan saturasi oksigen.

TABEL I
PENELITIAN TERDAHULU

| Authors | Input Parameters | Publisher | Predicted Output | Method | Accuracy |
|---|---|---|---|---|---|
| [6] | Breast cancer diagnosys data | Jurnal Informatika Vol.12 | Indentification of breast cancer disease | Logistic Regression | 98.4% |
| [8] | Age, sex, chest pain, trestbps, chol, fbs, restegc, thalac, exang, oldpeak, slope, ca, thul, target | Information System Development Volume 6 | Heart disease identification | Logistic Regression | 85.25% |
| [2] | Respiratory rate, oxygen saturation, and heart failure | Jurnal Ilmiah Permas: Jurnal Ilmiah STIKES Kendal Volume 10 | Relationship parameters to the heart failure patients | Medical Theory | Theory |
| [1] | Oxygen saturation and respiratory rate | Jurnal Ilmiah Wijaya Volume 12 | Relationship parameters to the head injury patients | Medical Theory | Theory |
| [5] | NEWS Parameters | BMJ Open Access | Relationship parameters to the cases in the hospital | Medical Theory | Theory |
| [9] | Demographic characteristic, race, ethnicity, vital signs, laboratory data, comordibities | Journal of Medical Internet Research | Predicting early respiratory failure | Logistic Regression | 91.5% |
| [10] | Breathing signal, oxygen saturation level | Cornell University Library | Detect apnea (respiratory disease) | Fully connected neural network | 99.88% |

Berdasarkan latar belakang yang sudah dijelaskan di atas, poin utama dalam penelitian ini adalah bagaimana persiapan data yang tepat agar menghasilkan hasil klasifikasi yang optimal dan bagaimana model dan hasil klasifikasi yang dihasilkan oleh model tersebut. Beberapa jurnal penelitian sebelumnya sudah ada yang menggunakan saturasi oksigen dan laju pernapasan sebagai parameter atau regresi logistik sebagai algoritma yang digunakan untuk melakukan klasifikasi [6] [8] [9] [10]. Namun penelitian-penelitian tersebut tidak menjelaskan secara rinci pengaruh variabel-variabel yang digunakan terhadap hasil model.

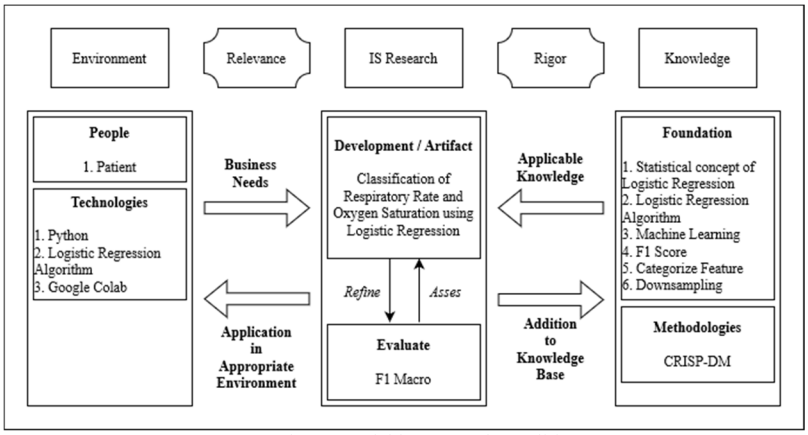
Tujuan dari penelitian ini adalah agar dapat memilih metode persiapan data yang tepat sehingga dapat menghasilkan klasifikasi kondisi pernapasan pasien yang optimal, serta mengetahui model yang dibangun dan mengetahui akurasi dari model klasifikasi tersebut dengan menggunakan regresi logistik. Diharapkan hasil penelitian ini dapat dikembangkan ke arah model prediksi dini kondisi pernapasan pasien oleh *Machine Learning* atau *Artificial Intelligence developer*. Sehingga masyarakat dapat melakukan pemeriksaan mandiri untuk mengetahui kondisi pernapasan. Bagi tenaga kesehatan dan peneliti di bidang kesehatan penelitian ini dapat menggunakan penelitian ini untuk menjelaskan analisis kesehatan dengan pendekatan data.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 8, No. 2, Juni 2023, Pp. 448-458

## II. METODE PENELITIAN

### A. *Model Konseptual*

Penelitian ini bertujuan untuk mengklasifikasikan kondisi pernapasan pasien berdasarkan parameter saturasi oksigen dan laju pernapasan menggunakan algoritma regresi logistik. Gambar 1 merupakan model konseptual yang digunakan dalam penelitian ini.



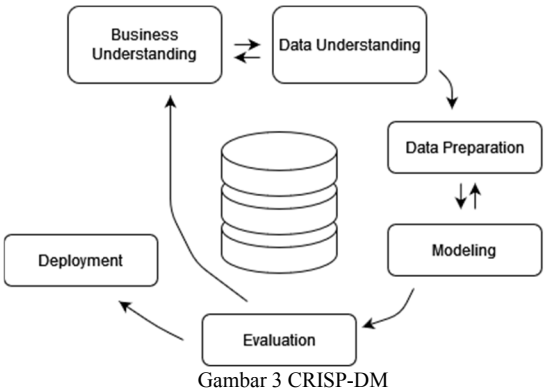

Gambar 2 Model konseptual penelitian

Model konseptual yang digambarkan pada gambar 2 menjelaskan keterkaitan aspek-aspek dalam penelitian ini. Pada model konseptual ini terdapat tiga elemen yang saling terkait, yaitu lingkungan, penelitian sistem informasi dan pengetahuan.

Penelitian yang dilakukan adalah klasifikasi laju pernapasan dan saturasi oksigen menggunakan regresi logistik yang nantinya model klasifikasi tersebut akan di evaluasi menggunakan F1-Macro. Lingkungan penelitian ini adalah pasien sebagai pemilik data yang digunakan dalam penelitian, python, Google Colab dan regresi logistik sebagai teknologi yang digunakan. Dasar pengetahuan yang digunakan dalam penelitian ini adalah konsep statistika regresi logistik, algoritma regresi logistik, machine learning, F1 Score, pengklasifikasian feature dan downsampling, sedangkan metodologi yang digunakan adalah CRISP-DM.

### B. *Sistematika Penyelesaian*

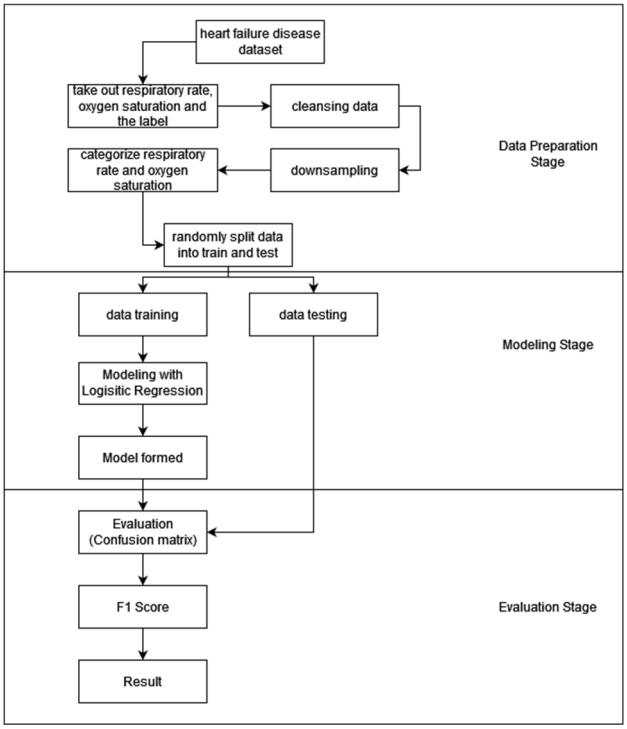
Dalam penelitian ini, peneliti menggunakan model *Cross Industry Standard Process for Data Mining* (CRISP DM) yang merupakan salah satu model proses data *mining* yang pada tahun 1966 dibangun oleh Integral Solution Ltd (ISL), Teradata, NCR Corporation, OHRA serta Daimler AG. Kemudian *framework* ini dikembangkan untuk dijadikan metodologi standar *non-proprietary* data *mining* oleh banyak perusahaan serta organisasi di daerah Eropa [11].




Gambar 3 CRISP-DM

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 8, No. 2, Juni 2023, Pp. 448-458

Terdapat 6 tahapan dalam CRISP DM yang menggambarkan siklus dari proyek seperti pada gambar 3. *Business understanding* merupakan tahap awal yang penting untuk mengetahui tujuan dari proyek, data apa yang akan digunakan dan bagaimana cara mendapatkan data tersebut. Data *understanding* merupakan tahapan memahami data sehingga dapat mengidentifikasi permasalahan yang akan dihadapi ketika menggunakan data tersebut.




Gambar 4 Tahapan pengerjaan penelitian

Kemudian pada tahap selanjutnya ada *data preparation* yang mana ditujukan untuk mengantisipasi masalah yang akan dihadapi ketika menggunakan data. Pada tahapan ini, akan sering ditinjau ulang ketika mendapati permasalahan dalam pengerjaan proyek. Dalam tahap ini juga, data akan dibagi menjadi dua bagian yaitu data *training* dan data *testing*. *Modeling* secara umum digunakan untuk membuat model prediktif ataupun model deskriptif menggunakan metode statistika ataupun *machine learning*. Tahap *evaluation* merupakan tahapan untuk menyesuaikan model yang sudah dibuat dengan tujuan awal dari proyek ini, untuk memastikan sudah sejalan dengan tujuan. Dan yang terakhir adalah *deployment*, yang merupakan tahap perencanaan penggunaan model.

Dari keseluruhan proses pada gambar 3, tahapan pengerjaan secara detail dari *data preparation* hingga tahapan *evaluation* dengan berdasarkan sistematika penyelesaian masalah yang sudah dijelaskan sebelumnya digambarkan pada gambar 4. Penelitian ini hanya akan membahas hingga tahap evaluasi. Hal ini dikarenakan untuk membangun model prediksi, khususnya prediksi untuk keperluan medis diperlukan data yang lebih banyak dan beragam agar pola dapat terlihat dengan lebih baik, sehingga perlu untuk dilakukannya *research* lebih mendalam. Maka dari itu, penelitian yang bertujuan untuk mengklasifikasikan kondisi pernapasan pasien ini hanya membahas hingga tahap evaluasi.

## C. *Pengumpulan Data*

Dalam penelitian ini terdapat dua variabel yang akan digunakan, yaitu saturasi oksigen dan laju pernapasan. Data yang digunakan merupakan data bersifat sekunder dari *website* penyedia *dataset* yaitu kaggle.com yang merupakan salah satu *platform* yang menyediakan *dataset* dan perlombaan *data science* dan *machine learning*[1].

TABEL II
SAMPEL DATA PENELITIAN

| No. | Respiratory Rate | Oxygen Saturation | Outcome |
|---|---|---|---|
| 1. | 15.652174 | 99.815789 | 1 |
| 2. | 21.75 | 94.384615 | 1 |
| 3. | 24.058824 | 96.117647 | 1 |
| 4. | 18.285714 | 96.964286 | 1 |
| 5. | 20.71875 | 99.73913 | 1 |

[1]*Dataset from Kaggle*: https://www.kaggle.com/datasets/saurabhshahane/in-hospital-mortality-prediction/code?select=data01.csv

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 8, No. 2, Juni 2023, Pp. 448-458

| | | | |
|---|---|---|---|
| 6. | 18.290323 | 98.166667 | 0 |
| 7. | 19.935484 | 98.043478 | 0 |
| 8. | 18.53125 | 97.8 | 0 |
| 9. | 17.936948 | 98.188513 | 0 |
| 10. | 17.129032 | 97.774194 | 0 |

Tabel II merupakan sampel data yang digunakan dalam penelitian ini. Data yang digunakan ini merupakan *dataset* dari penelitian mengenai prediksi kematian pada pasien gagal jantung di rumah sakit. *Dataset* tersebut terdiri dari 1177 data mentah dalam bentuk format CSV. Dari data tersebut, penelitian ini hanya mengambil kolom 'ID', '*respiratory rate'*, 'SP O2' dan '*outcome'* untuk menyesuaikan dengan tujuan penelitian. Data tersebut kemudian nantinya akan diolah menggunakan Python untuk diteliti, dimodelkan dan dianalisis.

### *D. Regresi Logistik*

Regresi logistik adalah salah satu model statistika pada *supervised learning*. Model ini digunakan untuk mengetahui efek dari variabel independen atau prediktor (X) pada variabel dependen atau respons (Y). Variabel dependen (Y) adalah nilai biner dengan kelas 0 dan 1 [8]. Data biner adalah bentuk paling umum dari data kategori, dan model ini merupakan model paling popular untuk menghadapi data biner. Untuk sebuah variabel respons Y, peluang variabel sukses dimodelkan dengan $P(Y = 1)$. Pada kasus satu variabel, variabel di lambangkan dengan $\pi(x)$ untuk menekankan bahwa nilai dari $P(Y = 1)$ bergantung pada nilai $x$ pada variabel tersebut. Model regresi logistik memiliki sebuah bentuk logit linear dari peluang sukses. Berikut aloritma dari peluang tersebut.

$$\log[\pi(x)] = \log\left[\frac{\pi(x)}{1 - \pi(x)}\right] = \alpha + \beta x \quad (1)$$

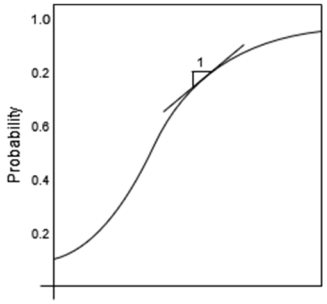
Untuk $x$, rumus menunjukkan $\pi(x)$ berubah sebagai sebuah fungsi dari S-*shaped* dari $x$. Regresi logistik memliki rumus untuk $\pi(x)$, menggunakan fungsi eksponensial, , $\exp(\alpha + \beta x) = e^{\alpha+\beta x}$ ,

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}} \quad (2)$$

Efek dari parameter $\beta$ menentukan nilai dari peningkatan atau penurunan dari kurva S-shaped untuk $\pi(x)$. $\beta$ mengidentifikasikan apakah kurva meningkat $(\beta > 0)$ atau menurun $(\beta < 0)$. Nilai dari perubahan akan meningkat ketika $|\beta|$ meningkat. Ketika $(\beta = 0)$, kurva akan mendatar membentuk sebuah garis horizontal yang lurus. Rumus regresi logistik menunjukkan logit akan meningkat dengan $\beta$ tiap peningkatan 1-unit pada nilai $x$. Dengan melakukan eksponensial pada kedua sisi persamaan regresi logistik, kita mendapatkan gambaran peluang dan rasio. Berikut peluang suksesnya:

$$\frac{\pi(x)}{1 - \pi(x)} = \exp(\alpha + \beta x) = e^{\alpha}(e^{\beta})^{x} \quad (3)$$

Inilah mengapa peluang dikalikan dengan $e^{\beta}$ untuk setiap peningkatan 1-unit pada nilai $x$, sehingga kurva s-shaped yang terbentuk nantinya akan terlihat seperti gambar 5.




Gambar 5 S-shaped untuk $\pi(x)$

### *E. Downsampling*

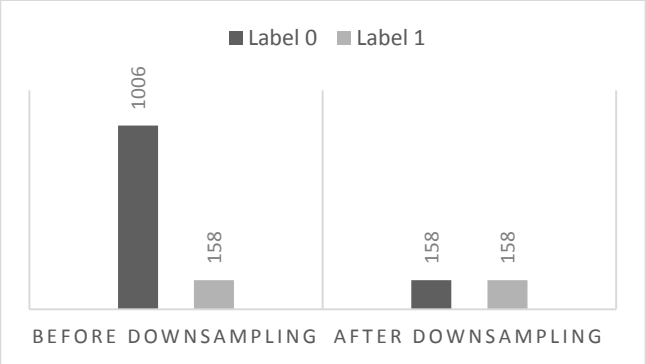
Pada akhir proses, terdapat 1164 data yang telah selesai di manipulasi dan siap untuk masuk ke tahap selanjutnya. Langkah selanjutnya, adalah mengetahui jumlah data dengan label atau target variabel 1 dan juga 0 untuk memastikan kondisi data apakah seimbang atau tidak. Karena *dataset* dengan kondisi variabel target tidak seimbang dapat menyebabkan hasil pemodelan tidak terklasifikasi dengan maksimal. Setelah dilakukan pengecekan pada tiap variabel target, ditemukan bahwa data yang akan digunakan untuk pemodelan merupakan data *imbalance* atau tidak

seimbang. Data dengan label atau variabel target 1 memiliki 158 baris data, sedangkan data dengan label 0 jauh lebih banyak dibanding dengan label 1 yaitu dengan 1006 baris data. Sehingga dapat dikatakan bahwa data yang digunakan ini memiliki variabel target 1 yang bersifat minoritas.

*Imbalance dataset* adalah kondisi di mana salah satu data klasifikasi memiliki jumlah data yang lebih kecil daripada kelompok klasifikasi lainnya. Kondisi ini berisiko menyebabkan model menjadi salah klasifikasi karena sebagian besar data kelompok minoritas lebih berharga daripada mayoritas [12]. *Dataset imbalance* memiliki tiga level: ringan, sedang, dan ekstrem. Tingkat ringan adalah ketika proporsi kelas minoritas adalah 20 - 40% dari *dataset*. Level sedang adalah ketika proporsi kelas minoritas adalah 1 - 20% dari *dataset*. Sedangkan level terakhir, ekstrem, adalah ketika kelas minoritas <1% dari *dataset*. Ada dua cara untuk menangani *dataset* yang tidak seimbang, Downsampling, dan Upweighting. Downsampling adalah teknik yang mengurangi jumlah kelas data mayoritas, sehingga kedua kelompok klasifikasi akan memiliki jumlah data yang seimbang. Upweighting adalah teknik yang menambahkan bobot pada kelas Downsampling [13]. Penelitian ini hanya akan menggunakan teknik Downsampling untuk menangani data yang tidak seimbang.

Kasus *imbalance* data yang terjadi pada data dalam penelitian ini ada pada tingkat *Moderate* dengan proporsi data minoritas sebesar 13.57% dari total jumlah data. Peneliti menggunakan metode Downsampling pada data kelompok mayoritas variabel target dengan memasang beberapa syarat ketika melakukan Downsampling tersebut, dikare-nakan teknik Downsampling dengan mengambil data secara random lebih berpengaruh jika kita memiliki banyak data [14]. Karena data mayoritas adalah data dari variabel target 0 atau 'aman', peneliti menentukan syarat berdasarkan syarat-syarat kondisi pernapasan pasien dikatakan aman dengan variabel SP $O_2$ dan laju pernapasan yang termasuk dalam rentang angka aman. Laju pernapasan yang memiliki rentang angka aman berkisar di antara angka 16 – 20, sedangkan saturasi oksigen yang aman berkisar pada angka lebih dari 95%, namun agar memudahkan mesin mempelajari data pasien yang memiliki kondisi pernapasan 'aman' untuk mendapatkan hasil prediksi terbaik, peneliti mengambil nilai saturasi oksigen $\geq$ 97%.

Setelah proses Downsampling, total data target variabel 0 ada sebanyak 188 data. Karena total data dengan variabel target 1 atau tidak aman terdapat 158 baris data, maka peneliti melakukan pengambilan sampel secara *random* sebanyak 158 data terhadap 188 data dengan variabel target 0 tersebut




Gambar 6 Grafik penerapan metode downsampling

Terlihat pada Gambar 6, akhir proses ini kedua kelompok target variabel memiliki jumlah masing-masing 158 baris data. Sehingga, total data yang digunakan dalam penelitian ini adalah sebanyak 316 baris data.

### *F. Mengategorikan Nilai pada Variabel*

Variabel *Respiratory rate* dan SP $O_2$ yang digunakan dalam penelitian ini memiliki data *raw* yang langsung berisi hasil pengukuran dari pasien tanpa dikategorikan dalam beberapa kelompok. Maka dari itu, untuk memaksimalkan hasil testing, peneliti mengelompokkan atau mengategorikan tiap nilai yang ada dalam beberapa kelompok berdasarkan batas-batas medis yang sudah ditentukan sebelumnya.

Variabel *Respiratory rate* akan dibagi menjadi 2 kategori yaitu 0 dan 1. Kedua kategori ini merepresentasikan 'rendah' dan 'tinggi' untuk kategori 1 dan 'normal' untuk kategori 0. Kategori rendah merupakan *Respiratory rate* yang memiliki nilai di bawah 16, dan kategori tinggi memiliki nilai *Respiratory rate* di atas 20. Untuk *Respiratory rate* yang dikategorikan dengan 0 adalah yang memiliki nilai dalam rentang 16 – 20. Sedangkan variabel SP $O_2$ atau saturasi oksigen juga akan dikategorikan menjadi 2 bagian, yaitu 0 merepresentasikan nilai SP $O_2$ diatas 97% dan 1 untuk merepresentasikan nilai yang tidak aman.

Pada Tabel III dapat terlihat hasil dari pengategorian nilai parameter *Respiratory rate* dengan nama variabel baru yaitu *new_respiratory_rate*, dan SP $O_2$ dengan nama variabel baru yaitu new_spo2. Kemudian kedua variabel baru ini yang akan digunakan dalam proses selanjutnya.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 8, No. 2, Juni 2023, Pp. 448-458

TABEL III
NILAI VARIABEL

| No. | Respiratory rate | new_respiratory_rate | SP $O_2$ | new_spo2 | outcome |
|---|---|---|---|---|---|
| 1. | 18.774194 | 0 | 99.478261 | 0 | 0 |
| 2. | 17.694444 | 0 | 99.742857 | 0 | 0 |
| 3. | 22.583333 | 1 | 94.607143 | 1 | 1 |
| 4. | 19.2 | 0 | 98.76 | 0 | 0 |
| 5. | 17.625 | 0 | 99.575 | 0 | 0 |
| 6. | 24.73913 | 1 | 98.26087 | 1 | 0 |
| 7. | 16.829268 | 0 | 99.230769 | 0 | 0 |

### G. *F1-Score*

Untuk melihat model regresi logistik pada penelitian ini lebih baik, kita membagi *dataset* ke dalam dua kelompok pada tahap persiapan data, *data training* dan *data testing*. Penelitian ini akan menggunakan *data training* untuk membangun model regresi logistik sedangkan *data testing* akan digunakan untuk melihat seberapa akurat model yang telah dibangun. Pemisahan data ini diambil dari data yang sudah siap untuk dimodelkan, dengan pembagian 805 atau 252 *data training* dan 20% atau 64 data untuk *data testing*. Metode F1-Score ini akan menggunakan data testing untuk mengevaluasi model yang telah di bangun.

Proses evaluasi standar menghitung proporsi prediksi yang benar untuk sampel yang digunakan sebagai *output* dan hasil prediksi yang diinginkan. Rumus yang digunakan untuk mengukur akurasi dengan N mewakili jumlah sampel:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} verdict \quad (4)$$

$$verdict = \begin{cases} 1, y_i = \hat{y}_i \\ 0, y_i \neq \hat{y}_i \end{cases} \quad (5)$$

Namun, metode evaluasi standar seperti di atas hanya menghasilkan evaluasi sederhana. Maka dari itu, F1-Sscore digunakan untuk menjelaskan hasil evaluasi lebih detail lagi. F1-Score adalah sebuah hasil dari proses perhitungan presisi dan *recall*. Untuk menemukan angka presisi dan *recall, y* adalah output yang diinginkan, sedangkan $\hat{y}$ adalah prediksi. Presisi didapatkan dengan membagi jumlah prediksi yang benar dengan total prediksi di sebuah kelas, dan *recall* didapatkan dengan membagi total prediksi yang benar dengan total *output* yang diinginkan dalam suatu kelas.

$$Precision = \frac{Total\ of\ correct\ prediction}{\|\hat{y}\|} \quad (6)$$

$$Recall = \frac{Total\ of\ correct\ prediction}{\|y\|} \quad (7)$$

$$F1 - Score = 2\,\frac{precision \times recall}{precision + recall} \quad (8)$$

Dalam menggunakan F1-Score, sebuah *confusion matrix* diperlukan untuk membantu menghitung menggunakan metode ini. *confusion matrix* akan mengembalikan nilai untuk $\|\hat{y}\|$ dan $\|y\|$ yang mana mempermudah perhitungan F1-Score. F1-Score yang didapatkan akan memberikan nilai berdasarkan kelas klasifikasi, dan untuk mendapatkan hasil secara keseluruhan kita dapat menggunakan tiga metode. Metode tersebut adalah F1-Macro, F1-Micro, dan F1-Weighted. Perbedaan dari ketiga metode ini adalah F1-Macro dan F1-Weighted akan menghitung dengan memberikan beban pada tiap kelas, sedangkan F1-Micro tidak. F1-Macro bekerja dengan merata-ratakan nilai F1-Score dari tiap kategori, yang mana diharapkan akan menghasilkan hasil akurasi yang lebih akurat [15].

## III. HASIL DAN PEMBAHASAN

### A. *Hasil*

Pada tahap akhir dari persiapan data, data akan dibagi dalam kelompok data independen dan data dependen pada masing-masing data *training* dan data testing. Hal ini bertujuan untuk memisahkan parameter-parameter tersebut dengan labelnya. Kemudian nantinya data ini akan dimasukkan ke model dan di prediksi ke dalam dua kategori, 0 untuk pasien dengan kondisi pernapasan pasien yang aman dan 1 untuk pasien dengan kondisi pernapasan yang

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 8, No. 2, Juni 2023, Pp. 448-458

tidak aman. Namun, sebelum dimasukkan ke dalam model, perlu untuk melakukan pengecekan terhadap data yang digunakan menggunakan nilai p-value untuk memastikan kelayakannya. Sebuah variabel dikatakan memenuhi syarat untuk dimasukkan ke dalam model ketika memiliki p-value di bawah 0.05 (uji wald dan likelihood) [16] dan memiliki nilai pseudo $r^2$ yang baik sebagai uji *goodness of fit* [17]. Tabel IV menampilkan nilai p-value dan koefisien tiap variabel serta nilai pseudo $r^2$ dari data yang digunakan pada penelitian ini:

TABEL IV
KOEFISIEN DAN P-VALUE VARIABEL INDEPENDEN

| | Coefficient | P-value |
|---|---|---|
| $\beta_0$ | 1.34384547 | |
| $X_1$ | 2.3713 | 0.000 |
| $X_2$ | 2.0209 | 0.000 |
| **LLR p-value (G)** | | 0.000 |

Pada table IV, Variabel $X_1$ adalah variabel Respiratory rate, $X_2$ adalah variabel SP $O_2$ atau saturasi oksigen dan $\beta_0$ adalah nilai *intercept* dari model yang digunakan. Uji Likelihood digunakan untuk menguji kelayakan model sehingga dapat diketahui apakah variabel independen yang digunakan pada penelitian ini benar-benar berpengaruh terhadap model. Pada tabel II di atas dapat terlihat, nilai Log-likelihood p-value model adalah 0.000 yang sudah memenuhi kriteria dari uji rasio likelihood, yaitu di bawah 0.05. Kemudian untuk melihat apakah masing-masing variabel mempengaruhi variabel dependen digunakan uji *wald,* untuk uji ini dapat terlihat pula pada tabel II bahwa kedua variabel yang digunakan dalam penelitian ini memiliki nilai p-value di bawah 0.05, tepatnya bernilai 0.00 pada masing-masing variabelnya. Sehingga kedua variabel penelitian ini dapat digunakan dalam melakukan pemodelan menggunakan regresi logistik, karena memenuhi kriteria uji Wald.

TABEL V
NILAI PSEUDO $R^2$

| | |
|---|---|
| **Pseudo-$R^2$** | 0.6909 |

Uji terakhir adalah uji *goodness of fit*. Berdasarkan tabel V, kedua variabel independen yang digunakan dalam penelitian ini 69% mempengaruhi variabel hasil atau variabel dependen. Hasil ini sudah cukup baik, namun juga mengindikasikan penelitian ini masih memerlukan tambahan variabel independen lainnya agar mendapatkan kecocokan yang lebih optimal.

Kemudian setelah melalui beberapa uji, model yang terbentuk jika proses pemodelan ini dijabarkan ke dalam rumus fungsi logistik akan berbentuk seperti ini:

$$\pi(x) = \frac{e^{\beta_0+\beta x}}{1+e^{\beta_0+\beta x}} \ldots(9)$$

$$\pi(x) = \frac{e^{1.34+2.37(x1)+(2.02)(x2)}}{1+e^{1.34+2.37(x1)+(2.02)(x2)}} \ldots(10)$$

$\pi(x)$ adalah nilai peluang dari satu kejadian dengan dua kemungkinan, 0 dan 1. Nilai $\alpha$ di peroleh dari nilai koefisien *intercept* model, sedangkan nilai $\beta$ diperoleh dari nilai koefisien dari masing-masing variabel $X_1$ dan $X_2$, yang secara berurutan adalah variabel *Respiratory rate* dan SP O2. Sedangkan untuk nilai $x$ sendiri merupakan nilai dari variabel *Respiratory rate* dan SP O2.

Regresi logistik yang digunakan dipanggil menggunakan fungsi dari *library* sklearn model linear. Algoritma ini hanya akan menggunakan variabel independen dan memprediksi klasifikasinya menggunakan variabel-variabel tersebut. Setelah mendapatkan prediksi klasifikasi, model akan menghitung akurasi. Akurasi ini dihitung dengan membandingkan variabel dependen atau hasil prediksi klasifikasi oleh model dengan variabel dependen yang diinginkan. Hasil akurasi menggunakan data *training* dan data *testing* masing-masing adalah 95% dan 87%.

Penelitian ini menggunakan F1-Macro untuk mendapatkan evaluasi data yang lebih detail untuk mengevaluasi model ini. Evaluasi model ini menggunakan nilai presisi dan *recall* untuk mendapatkan nilai F1-Macro yang mana menggunakan data testing. Seperti yang sudah dijelaskan sebelumnya, metode evaluasi ini menggunakan *confusion matrix* agar perhitungannya lebih mudah dilakukan.

TABEL VI
CONFUSION MATRIX

| | Actual Positive | Actual Negative |
|---|---|---|
| **Predicted Positive** | 32 | 0 |
| **Predicted Negative** | 8 | 24 |

Dengan menggunakan hasil matriks pada tabel VI, dapat dihitung nilai presisi, *recall*, F1-Macro, dan F1-Score untuk mengevaluasi model ini.

TABEL VII
EVALUASI F1-SCORE

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| **0** | 0.80 | 1.00 | 0.89 | 32 |
| **1** | 1.00 | 0.75 | 0.86 | 32 |
| **Accuracy** | | | 0.88 | 64 |
| **Macro Avg** | 0.90 | 0.88 | 0.87 | 64 |

Tabel VII menampilkan nilai detail dari presisi, *recall*, F1-Macro, dan F1-Score. Nilai F1-Score secara umum adalah 88% dan dengan menggunakan metode F1-Macro adalah 87%. Evaluasi model regresi logistik menggunakan F1-Macro ini tergolong cukup baik untuk sebuah model.

TABEL VIII
STUDI PERBANDINGAN MENGGUNAKAN REGRESI LOGISTIK

| Input Data | Accuracy Rate (%) | Ref. |
|---|---|---|
| Oxygen saturation, respiratory rate | 87% | This work |
| Age, sex, chest pain, trestbps, chol, fbs, restegc, thalac, ex-ang, oldpeak, slope, ca, thul, target | 85.25% | [8] |
| Breast cancer diag-nosys data | 98.4% | [6] |
| Demographic char-acteristic, race, eth-nicity, vital signs, la-boratory data, com-ordibities | 91.5% | [9] |

Tabel VIII menunjukkan studi perbandingan antara penelitian yang sebelumnya pernah dilakukan menggunakan regresi logistik namun menggunakan data *input* yang berbeda. Dapat terlihat bahwa beberapa penelitian terdahulu belum ada yang hanya menggunakan saturasi oksigen dan laju pernapasan sebagai parameter dengan algoritma regresi logistik.

## *B. Pembahasan*

Hasil klasifikasi kondisi pernapasan pasien memiliki hasil akurasi *training* dan evaluasi yang baik, yaitu 95% dan 87%. Karna pembangunan model klasifikasi ini menggunakan algoritma regresi logistik, maka perlu untuk memeriksa pengaruh kedua variabel, yaitu laju pernapasan (*Respiratory rate*) dan saturasi oksigen (SP O2) pada model. Pada Langkah mengategorikan nilai variabel, peneliti mengategorikan Y (*outcome*), *Respiratory rate* ($X_1$) dan saturasi oksigen ($X_2$) dengan setara, yaitu kondisi aman di labelkan oleh 0 dan tidak aman oleh 1.

Ketika nilai $X_1$ dan $X_2$ adalah sama, maka dampak pada Y atau variabel dependen adalah senilai angka koefisien *intercept*. Sedangkan koefisien yang dimiliki oleh $X_1$ dan $X_2$ adalah untuk mengukur kontribusi oleh kedua variabel tersebut terhadap Y. Untuk menginterpretasikan nilai koefisien, perlu menggunakan *odd ratio* atau rasio peluang pada tiap variabel independen. *Odd ratio* sendiri adalah pengukuran yang dilakukan untuk mengukur kekuatan kejadian. Koefisien yang memiliki nilai positif bermakna bahwa setiap kenaikan satu nilai, akan meningkatkan peluang Y oleh nilai koefisien variabel tersebut [18].

Rasio peluang dihitung dengan menggunakan fungsi eksponensial bersama dengan koefisien dari variabel independen. Sehingga rasio peluang dari *Respiratory rate* ($X_1$) bentuknya seperti ini: $e^{2.37\ 13} = 10.7113$, yang mana angkat tersebut datang dari nilai koefisien $X_1$. Interpretasi dari hasil rasio peluang adalah pasien dengan laju pernapasan yang abnormal memiliki risiko mengalami kondisi pernapasan yang tidak aman 10.7 kali lebih besar dibandingkan dengan pasien yang memiliki laju pernapasan normal. Sedangkan rasio peluang dari saturasi oksigen ($X_2$) berbentuk seperti ini: $e^{2.0209} = 7.5451$, yang mana angka tersebut berasal dari nilai koefisien $X_2$. Intepretasi dari hasil rasio peluang ini adalah pasien dengan saturasi oksigen abnormal memiliki resiko mengalami kondisi pernapasan tidak aman 7.5 kali lebih besar dari pada pasien dengan saturasi oksigen yang normal.

Interpretasi dari variabel di atas sejalan dengan teori medis saturasi oksigen dan laju pernapasan. Pada teori medis, pasien dengan laju pernapasan dan saturasi oksigen yang abnormal memiliki peluang untuk terindikasi memiliki permasalahan pada pernapasannya. Namun terdapat ketimpangan antara hasil interpretasi dengan teori medis untuk setiap variabelnya. Karena berdasarkan teori medis, seharusnya pasien dengan saturasi oksigen abnormal memiliki peluang lebih besar mengalami kondisi pernapasan tidak aman dibandingkan dengan pasien dengan laju pernapasan

yang tidak normal. Hal ini dikarenakan laju pernapasan memiliki lebih banyak faktor yang menyebabkannya tidak normal, seperti umur, aktivitas, emosi dan lainnya yang sebenarnya tidak berkaitan dengan kondisi pernapasan yang tidak aman.

Ketimpangan ini juga dapat disebabkan oleh beberapa hal lainnya, termasuk kurangnya variabel independen atau kurangnya data yang digunakan dalam penelitian. Oleh karena itu, untuk penelitian ke depan atau pengembangan penelitian ini, perlu untuk menambahkan variabel independen lainnya dan meningkatkan jumlah data, sehingga dapat membangun data yang lebih akurat.

## IV. KESIMPULAN

Penelitian ini menggunakan algoritma regresi logistik untuk membangun model klasifikasi kondisi pernapasan pasien dengan saturasi oksigen dan laju pernapasan sebagai parameter. Klasifikasi tersebut diharapkan dapat menjadi prediksi awal kondisi pernapasan pasien dan selanjutnya berkembang menjadi prediksi awal penyakit pernapasan.

Dari hasil dan proses evaluasi, model klasifikasi ini memiliki nilai akurasi yang baik, yaitu 87% dengan menggunakan 316 total data. Hasil penelitian ini juga mengikuti teori saturasi oksigen dan laju pernapasan medis yang ada. Namun, ada perbedaan mengenai hasil prediksi dan teori medis mengenai risiko pasien dengan tingkat pernapasan dan saturasi oksigen yang tidak normal. Ke depannya, penelitian ini akan lebih baik jika menggunakan data medis yang lebih nyata dan menambahkan variabel independen, sehingga mesin dapat belajar lebih banyak dari data dan melanjutkan proses penelitian hingga tahap akhir, tahap penyebaran.

## DAFTAR PUSTAKA

[1] C. Sumiarty and F. A. Sulistyo, "HUBUNGAN RESPIRATORY RATE (RR) DENGAN OXYGEN SATURATION (SpO2) PADA PASIEN CEDERA KEPALA," *Jurnal Ilmiah Wijaya,* pp. 101-109, 2020.

[2] Isrofah, A. Indriono and I. Mushafiyah, "TIDUR DAN SATURASI OKSIGEN PADA PASIEN CONGESTIVE HEARTH FAILLURE," *Jurnal Ilmiah Permas: Jurnal Ilmiah STIKES Kendal,* p. 557–568, 2020.

[3] Website Pemerintah Daerah Jakarta, "corona.jakarta.go.id," 19 Januari 2022. [Online]. Available: https://corona.jakarta.go.id/id/artikel/varian-varian-covid-19-apa-perbedaannya.

[4] M. Ilham, I. Sarwili and S. Kamilah, "Prone Position Dapat Meningkatkan Kadar Saturasi Oksigen Pada Pasien Covid-19," *Open access Jakarta Journal of health sciences,* pp. 146-152, 2022.

[5] L. Chen, H. Zheng, L. Chen, S. Wu and S. Wang, "National Early Warning Score in Predicting Severe Adverse Outcomes of Emergency Medicine Patients: A Retrospective Cohort Study," *Journal of Multidisciplinary Healthcare,* pp. 2067-2078, 2021.

[6] R. A. Vinarti and W. Anggraeni, "Identifikasi Faktor Prediksi Diagnosis Tingkat Keganasan Kanker Payudara Metode Stepwise Binary Logistic Regression," *Jurnal Informatika,* pp. 70-76, 2014.

[7] A. Agresti, An Introduction to Categorical Data Analysis, United States of America: John Wiley & Sons, Inc., 2019.

[8] J. J. Pangaribuan, H. Tanjaya and Kenichi, "MENDETEKSI PENYAKIT JANTUNG MENGGUNAKAN MACHINE LEARNING DENGAN ALGORITMA LOGISTIC REGRESSION," *INFORMATION SYSTEM DEVELOPMENT,* 2021.

[9] M. Siavash Bolourani, M. P. Max Brenner, M. Ping Wang, M. M. Thomas McGinn, M. Jamie S Hirsch, M. Douglas Barnaby and T. P. P. M. B. Zanos, "A Machine Learning Prediction Model of Respiratory Failure Within 48 Hours of Patient Admission for COVID-19: Model Development and Validation," *Journal of Medical Internet Research,* vol. 23, no. 2, 2021.

[10] O. Hassan, S. Shamsir and S. K. Islam, "Machine Learning Based Hardware Model for a Biomedical System for Prediction of Respiratory Failure," in *020 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, 2021.

[11] F. B. Tuga Mauritsius, "mmsi.binus.ac.id," 18 September 2020. [Online]. Available: https://mmsi.binus.ac.id/2020/09/18/cross-industry-standard-process-for-data-mining-crisp-dm/.

[12] P. Dr.D.Ramyachitra, "IMBALANCED DATASET CLASSIFICATION AND SOLUTIONS: A REVIEW," *International Journal of Computing and Business Research (IJCBR) ,* vol. 5, no. 4 July 2014, 2014.

[13] Google Developers Courses, "developers.google.com," 1 November 2021. [Online]. Available: https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data.

[14] J. R. M. A. Roweida Mohammed, "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results," in *11th International Conference on Information and Communication Systems (ICICS)*, 2020.

[15] J. W. G. Putra, Pengenalan Konsep Pembelajaran Mesin dan Deep Learning, Tokyo, 2020.

[16] N. Rajagukguk, D. Ispriyanti and Y. Wilandari, "Perbandingan Metode Klasifikasi Regresi Logistik Biner dan Naive Bayes pada Status Pengguna KB di Kota Tegal Tahun 2014," *Jurnal Gaussian,* pp. 365 - 374, 2015.

[17] G. A. J. Hemmert, L. M. Schons, J. Wieseke and H. Schimmelpfennig, "Log-likelihood-based Pseudo-R^2 in Logistic Regression: Deriving Sample-sensitive Benchmarks," *Sociological Methods & Research,* pp. 1-25, 2016.

[18] statology, "statology.org," 20 March 2019. [Online]. Available: https://www.statology.org/read-interpret-regression-table/.