

**PREDIKSI PENYAKIT PARU-PARU BERDASARKAN GAYA HIDUP
DAN PENYAKIT BAWAAN MENGGUNAKAN ALGORITMA LOGISTIC
REGRESSION**

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi S1 Informatika



disusun oleh
YOGA FIRMANSYAH
21.11.4529

Kepada

FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2025

**PREDIKSI PENYAKIT PARU-PARU BERDASARKAN GAYA HIDUP
DAN PENYAKIT BAWAAN MENGGUNAKAN ALGORITMA LOGISTIC
REGRESSION**

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi S1 Informatika



disusun oleh

YOGA FIRMANSYAH

21.11.4529

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2025

HALAMAN PERSETUJUAN

SKRIPSI

**PREDIKSI PENYAKIT PARU-PARU BERDASARKAN GAYA HIDUP
DAN PENYAKIT BAWAAN MENGGUNAKAN ALGORITMA LOGISTIC
REGRESSION**

yang disusun dan diajukan oleh

Yoga Firmansyah

21.11.4529

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 28 Mei 2025

Dosen Pembimbing,



Rumini, S.Kom., M.Kom.

NIK. 190302246

HALAMAN PENGESAHAN

SKRIPSI

**PREDIKSI PENYAKIT PARU-PARU BERDASARKAN GAYA HIDUP
DAN PENYAKIT BAWAAN MENGGUNAKAN ALGORITMA LOGISTIC
REGRESSION**

yang disusun dan diajukan oleh

Yoga Firmansyah

21.11.4529

Telah dipertahankan di depan Dewan Penguji
pada tanggal 28 Mei 2025

Susunan Dewan Penguji

Nama Penguji

Tanda Tangan

Ninik Tri Hartanti, S.Kom., M.Kom.,
NIK. 190302330

Senle Destya, S.T., M.Kom.,
NIK. 190302312

Rumini, S.Kom., M.Kom.,
NIK. 190302246

Skrripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 28 Mei 2025

DEKAN FAKULTAS ILMU KOMPUTER



Prof. Dr. Kusriani, M.Kom.,

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Yoga Firmasyah
NIM : 21.11.4529

Menyatakan bahwa Skripsi dengan judul berikut:

PREDIKSI PENYAKIT PARU-PARU BERDASARKAN GAYA HIDUP DAN PENYAKIT BAWAAN MENGGUNAKAN ALGORITMA LOGISTIC REGRESSION

Dosen Pembimbing : Rumini, M.Kom

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 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, 28 Mei 2025

Yang Menyatakan,



Yoga Firmansyah

HALAMAN PERSEMBAHAN

Segala puji dan syukur penulis panjatkan ke hadirat Allah SWT atas limpahan rahmat dan hidayah-Nya, sehingga penulis dapat menyelesaikan penyusunan skripsi ini. Shalawat dan salam semoga senantiasa tercurah kepada Nabi Muhammad SAW sebagai suri teladan umat. Dengan penuh rasa syukur dan kerendahan hati, penulis menyadari bahwa pencapaian ini tidak diperoleh tanpa melalui berbagai proses dan tantangan. Berkat ketulusan niat, doa yang senantiasa terpanjat, serta dukungan dari berbagai pihak, khususnya keluarga dan orang-orang terdekat, penulisan skripsi ini dapat terselesaikan dengan baik. Oleh karena itu, dengan penuh penghargaan dan hormat, karya ini penulis persembahkan kepada:

1. Penulis menyampaikan penghargaan dan rasa terima kasih yang tulus kepada Bapak Ferdinan Marcos dan Ibu Suyaini, atas doa yang senantiasa dipanjatkan, kasih sayang yang tulus, serta dukungan yang tiada henti. Bapak dan Ibu merupakan sumber kekuatan utama yang mendorong penulis untuk terus berjuang dan menyelesaikan setiap proses serta tantangan dalam penulisan skripsi ini. Pencapaian ini tidak akan mungkin terwujud tanpa kehadiran, cinta, dan pengorbanan yang telah Bapak dan Ibu berikan sepanjang perjalanan ini.
2. Kakek dan nenek tercinta Samin, Mulyani, Nur Husna serta almarhum kakek Junaidi Saat yang telah lebih dahulu berpulang, penulis menyampaikan terima kasih atas doa dan semangat yang senantiasa diberikan. Kehadiran serta doa tulus dari kalian menjadi sumber kekuatan yang sangat berarti bagi penulis dalam menyelesaikan perjalanan ini.
3. Kakak tercinta, Wahyudi dan Fernando Eran Juliandi, penulis mengucapkan terima kasih atas segala bentuk dukungan, perhatian, dan kasih sayang yang senantiasa menyertai di setiap langkah. Kehadiran dan kebersamaan kalian menjadi sumber semangat bagi penulis untuk terus melangkah maju dan menyelesaikan proses ini dengan penuh keyakinan.

4. Adik tercinta, Muhammad Altaf Fauzan, kehadiranmu merupakan anugerah yang berharga bagi keluarga serta menjadi sumber semangat yang menguatkan hati penulis dalam menjalani setiap proses.
5. Penulis juga menyampaikan rasa terima kasih yang mendalam kepada Luri Pijria Diningsih, yang senantiasa menjadi pendengar setia dalam setiap keluh kesah selama proses penulisan skripsi ini. Dukungan, motivasi, pengingat, serta pemahaman yang tulus yang diberikan telah menjadi salah satu kekuatan penting yang membantu penulis menyelesaikan skripsi ini dengan baik.
6. Penulis juga ingin mengucapkan terima kasih yang sebesar-besarnya kepada teman-teman seperjuangan, yang selalu menemani dan berbagi dalam setiap suka maupun duka. Kebersamaan, semangat, dan dukungan yang tak ternilai sepanjang perjalanan ini sangat berarti bagi penulis dan menjadi bagian penting dalam menyelesaikan tugas ini.
7. Penulis juga mempersembahkan karya ini kepada almamater tercinta, dengan harapan semoga karya yang sederhana ini dapat memberikan sumbangsih yang bermanfaat bagi perkembangan ilmu pengetahuan dan kemajuan almamater, serta memberi manfaat yang lebih luas bagi masyarakat.

Dengan penuh harapan, semoga skripsi ini tidak hanya menjadi bukti dari perjuangan, tetapi juga dapat memberikan manfaat serta menjadi langkah awal menuju masa depan yang lebih gemilang.

KATA PENGANTAR

Dengan mengucapkan puji syukur kehadiran Allah SWT yang telah melimpahkan Rahmat dan karunia-Nya, sehingga pada akhirnya penulis dapat menyelesaikan skripsi ini yang berjudul **“Prediksi Penyakit Paru-Paru Berdasarkan Gaya Hidup dan Penyakit Bawaan Menggunakan Algoritma Logistic regression”**. Penulisan skripsi ini dibuat sebagai salah satu syarat kelulusan program Sarjana Program Studi Informatika Fakultas Ilmu Komputer Universitas Amikom Yogyakarta. Penulis menyadari bahwa tanpa bimbingan dan dorongan dari semua pihak, maka penulisan skripsi ini tidak akan berjalan lancar. Oleh karena itu, pada kesempatan ini, izinkalah penulis menyampaikan ucapan terima kasih kepada:

1. Bapak Prof. Dr. M. Suyanto, MM., selaku Rektor Universitas Amikom Yogyakarta.
2. Ibu Prof. Dr. Kusriani, S.Kom., M.Kom., selaku Dekan Fakultas Ilmu Komputer Universitas Amikom Yogyakarta.
3. Ibu Eli Pujastuti, S.Kom., M.Kom., selaku Ketua Program Studi Informatika Universitas Amikom Yogyakarta.
4. Ibu Rumini, S.Kom., M.Kom., selaku dosen wali serta Dosen Pembimbing Skripsi yang telah menyediakan waktu, tenaga, dan pikiran untuk mengarahkan penulis dalam penyusunan skripsi ini. Terima kasih atas ilmu dan pengetahuan yang telah diberikan.
5. Seluruh Staf Pengajar dan Pegawai Fakultas Ilmu Komputer Universitas Amikom Yogyakarta yang telah memberikan bantuan dan dukungan kepada penulis selama menjalani studi di Universitas Amikom Yogyakarta. Semua kontribusi kalian sangat berarti bagi penulis dalam meraih keberhasilan ini.
6. Penulis menyampaikan rasa terima kasih yang sebesar-besarnya kepada keluarga tercinta Bapak, Ibu, Kakak, Adik, serta seluruh anggota keluarga atas doa yang tiada henti, kasih sayang yang tulus, dan dukungan yang senantiasa diberikan sepanjang proses penulisan skripsi ini. Kehadiran dan cinta dari keluarga menjadi sumber kekuatan dan motivasi yang luar biasa, yang

mendorong penulis untuk terus berusaha, melewati berbagai tantangan, dan menyelesaikan skripsi ini dengan sebaik-baiknya. Tanpa dukungan dan pengorbanan kalian, pencapaian ini tentu tidak akan dapat terwujud.

Penulis menyadari bahwa penyusunan skripsi ini masih jauh dari sempurna. Oleh karena itu, kritik dan saran yang membangun sangat diharapkan guna perbaikan di masa yang akan datang. Semoga karya ini dapat memberikan manfaat bagi dunia akademik serta menjadi kontribusi sederhana dalam pengembangan ilmu pengetahuan, khususnya di bidang prediksi dan *machine learning*.

Yogyakarta, 28 Mei 2025



Penulis

DAFTAR ISI

HALAMAN JUDUL	i
HALAMAN PERSETUJUAN	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERSEMBAHAN	v
KATA PENGANTAR	vii
DAFTAR ISI	ix
DAFTAR TABEL	xiii
DAFTAR GAMBAR	xiv
DAFTAR LAMPIRAN	xvi
DAFTAR LAMBANG DAN SINGKATAN	xvii
DAFTAR ISTILAH	xviii
INTISARI	xix
ABSTRACT	xx
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	2
1.3 Batasan Masalah	2
1.4 Tujuan Penelitian	4
1.5 Manfaat Penelitian	5
1.6 Sistematika Penulisan	5
BAB II TINJAUAN PUSTAKA	6
2.1 Studi Literatur	6
2.2 Dasar Teori	16
2.2.1 Data Mining	16

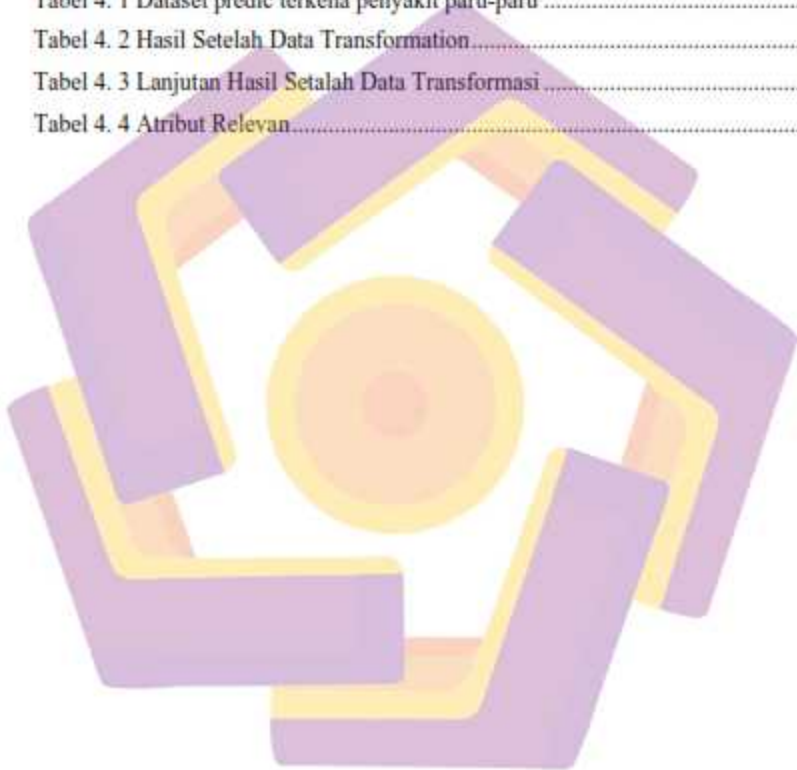
2.2.2 <i>Logistic regression</i> (Regresi Logistik).....	16
2.2.3 <i>Exploratory Data Analysis (EDA)</i>	18
2.2.4 <i>Preprocessing</i>	18
2.2.5 <i>Data Splitting</i> (Pemisahan Data).....	19
2.2.6 <i>Prediction</i> (Prediksi)	19
2.2.7 <i>Confusion Matrix</i>	20
2.2.8 <i>Accuracy</i>	21
2.2.9 <i>Precision</i>	21
2.2.10 <i>Recall</i>	22
2.2.11 <i>F1-score</i>	22
2.2.12 <i>ROC Curve</i>	23
2.2.13 <i>Cross Validation</i>	23
2.2.14 <i>Best Hyperparameter</i>	23
2.2.15 <i>Python</i>	23
2.2.16 <i>Google Colab</i>	24
2.2.17 <i>Streamlit</i>	24
BAB III METODE PENELITIAN	25
3.1 <i>Objek Penelitian</i>	25
3.2 <i>Alur Penelitian</i>	25
3.2.1 <i>Identifikasi Masalah</i>	26
3.2.2 <i>Pengumpulan Data</i>	26
3.2.3 <i>Exploratory Data Analysis (EDA)</i>	28
3.2.4 <i>Preprocessing</i>	29
3.2.5 <i>Modeling</i>	31
3.2.6 <i>Implementasi Perhitungan Logistic regression</i>	31

3.2.7 <i>Evaluation</i>	32
3.3 <i>Alat dan Bahan</i>	33
3.3.1 <i>Alat</i>	33
3.3.2 <i>Bahan</i>	34
BAB IV HASIL DAN PEMBAHASAN	35
4.1 <i>Sumber Data</i>	35
4.2 <i>Exploratory Data Analysis (EDA)</i>	36
4.2.1 <i>Distribusi Data</i>	36
4.2.2 <i>Korelasi Hubungan Antar Variabel</i>	36
4.2.3 <i>Menampilkan 5 Baris Pertama</i>	38
4.2.4 <i>Menampilkan Informasi <i>Dataframe</i></i>	38
4.2.5 <i>Menampilkan Statistik Deskriptif Dataset</i>	39
4.2.6 <i>Mengecek Nilai Hilang (<i>Missing Value</i>)</i>	40
4.3 <i>Preprocessing</i>	40
4.3.1 <i>Data Cleaning</i>	40
4.3.2 <i>Data Duplication</i>	41
4.3.3 <i>Data Transformation</i>	42
4.3.4 <i>Feature Selection</i>	43
4.3.5 <i>Splitting Data</i>	43
4.3.6 <i>Feature Scaling (Normalisasi/Standarisasi Data)</i>	44
4.3.7 <i>PCA</i>	44
4.4 <i>Data Mining/Modeling</i>	46
4.4.1 <i>Model Training</i>	46
4.4.2 <i>Model Preditions</i>	47
4.5 <i>Evaluation</i>	47

4.5.1	<i>Cross Validation</i>	47
4.5.2	<i>Best Hyperparameters</i>	48
4.5.3	<i>Classification Report</i>	49
4.5.4	<i>Confusion Matrix</i>	51
4.5.5	<i>ROC-AUC</i>	53
4.6	<i>Feature Importance</i>	54
4.7	<i>Save Model</i>	55
4.8	<i>Deployment Streamlit</i>	56
4.8.1	Membuat Modul <i>Library Sklearn Streamlit</i>	56
4.8.2	Membuat Aplikasi Prediksi Penyakit Paru-Paru Berbasis <i>Web</i>	57
BAB V PENUTUP		62
5.1	Kesimpulan	62
5.2	Saran	62
REFERENSI		64
LAMPIRAN		68

DAFTAR TABEL

Tabel 2. 1 Keaslian Penelitian	11
Tabel 2. 2 Confusion Matrix	20
Tabel 4. 1 Dataset predic terkena penyakit paru-paru	35
Tabel 4. 2 Hasil Setelah Data Transformation	42
Tabel 4. 3 Lanjutan Hasil Setelah Data Transformasi	43
Tabel 4. 4 Atribut Relevan	43



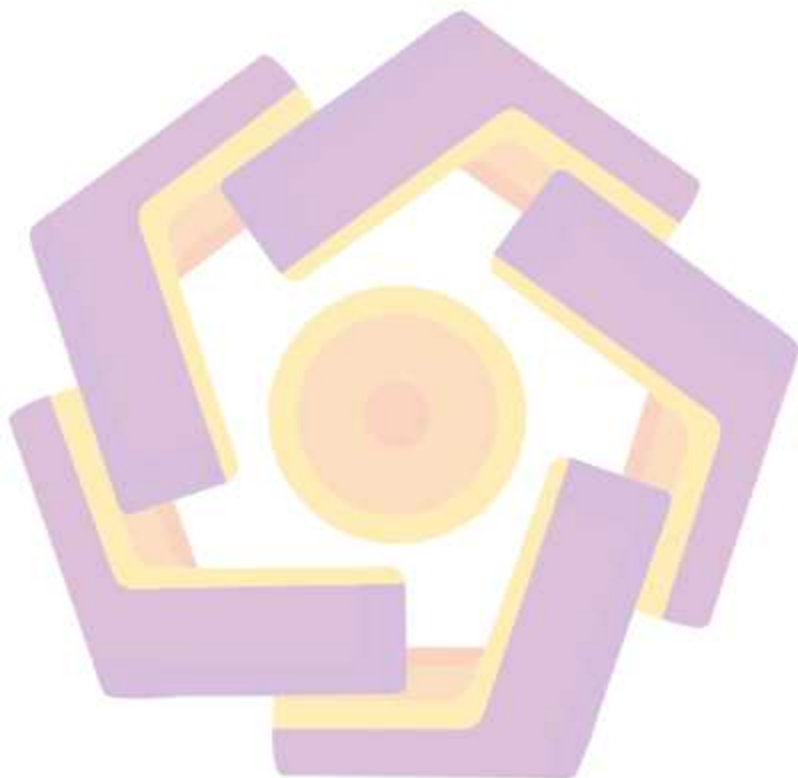
DAFTAR GAMBAR

Gambar 3. 1 Alur Penelitian	25
Gambar 3. 2 Dataset prediksi penyakit paru-paru	26
Gambar 4. 1 Boxplot Distribusi Data.....	36
Gambar 4. 2 Korelasi Hubungan Antar Variabel.....	37
Gambar 4. 3 Source Code menampilkan 5 baris pertama	38
Gambar 4. 4 Menampilkan 5 baris pertama	38
Gambar 4. 5 Source Code informasi dataframe	38
Gambar 4. 6 Informasi dataframe	39
Gambar 4. 7 Source Code statistik deskriptif dataframe	39
Gambar 4. 8 Statistik deskriptif dataset	39
Gambar 4. 9 Source Code Mengecek Nilai Hilang	40
Gambar 4. 10 Hasil mengecek nilai hilang	40
Gambar 4. 11 Source Code Data Cleaning	40
Gambar 4. 12 Hasil Data Cleaning	41
Gambar 4. 13 Data Duplication	41
Gambar 4. 14 Source Code Data Transformation.....	42
Gambar 4. 15 Hasil Sebelum Data Transformation.....	42
Gambar 4. 16 Source Code Splitting Data	44
Gambar 4. 17 Source Code Feature Scaling	44
Gambar 4. 18 Source Code PCA	45
Gambar 4. 19 Visualisasi PCA (2 komponen utama)	45
Gambar 4. 20 Model Training	46
Gambar 4. 21 Model Predictions	47
Gambar 4. 22 Source Code Cross Validation	47
Gambar 4. 23 Cross Validation.....	48
Gambar 4. 24 <i>Best Hyperparameters</i>	48
Gambar 4. 25 Classification Report	49
Gambar 4. 26 Visualiasasi Classification Report	50
Gambar 4. 27 Confusion Matrix	51

Gambar 4. 28 Visualisasi Confusion Matrix.....	52
Gambar 4. 29 ROC CURVE	53
Gambar 4. 30 Visualiasasi ROC CURVE.....	53
Gambar 4. 31 Source code Feature Importance.....	54
Gambar 4. 32 Visualisasi Feature Impotance	55
Gambar 4. 33 Save Model.....	55
Gambar 4. 34 Deployment Streamlit	56
Gambar 4. 35 Import modul untuk membuat aplikasi	57
Gambar 4. 36 Memasukkan file model prediksi ke program.....	57
Gambar 4. 37 Membuat inputan data ke aplikasi.....	57
Gambar 4. 38 Membuat inputan data Encoding	58
Gambar 4. 39 Membuat inputan Tombol Prediksi.....	58
Gambar 4. 40 Menampilkan input basis web.....	59
Gambar 4. 41 Hasil jika terkena penyakit paru-paru	60
Gambar 4. 42 Hasil jika tidak terkena penyakit paru-paru	61

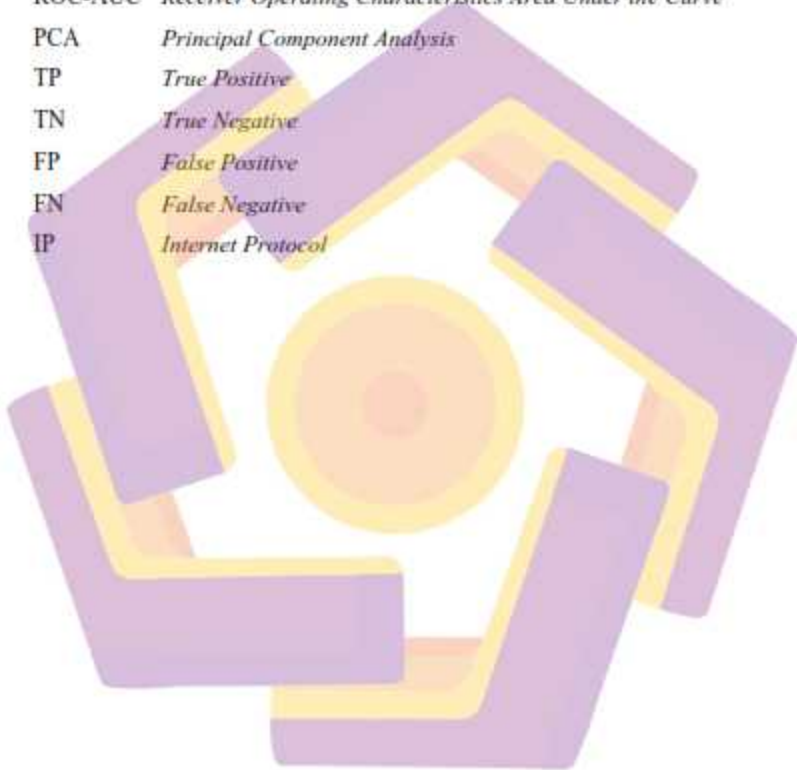
DAFTAR LAMPIRAN

Lampiran 1 Dataset prediksi penyakit paru-paru.....	68
Lampiran 2 Source code penelitian.....	68
Lampiran 3 Link Local Host Deployment Streamlit	68



DAFTAR LAMBANG DAN SINGKATAN

β	Nilai Beta
URL	<i>Uniform Resource Locator</i>
EDA	<i>Exploratory Data Analysis</i>
ROC-AUC	<i>Receiver Operating Characteristics Area Under the Curve</i>
PCA	<i>Principal Component Analysis</i>
TP	<i>True Positive</i>
TN	<i>True Negative</i>
FP	<i>False Positive</i>
FN	<i>False Negative</i>
IP	<i>Internet Protocol</i>



DAFTAR ISTILAH

<i>Prediction</i>	Prediksi Penyakit Paru-Paru.
<i>Missing Value</i>	Data yang tidak tersedia atau hilang.
<i>Overfitting</i>	Skor akurasi rendah pada data testing, tetapi skor tinggi pada data training.
<i>C</i>	<i>Best Hyperparameter</i> yang mengatur keseimbangan antara margin dan kesalahan klasifikasi.
<i>Confusion Matrix</i>	Sebuah tabel yang digunakan untuk mengevaluasi kinerja model klasifikasi.
<i>Accuracy</i>	Persentase prediksi yang bertujuan untuk mengukur seberapa sering model membuat prediksi yang benar.
<i>Precision</i>	Berapa banyak dari semua prediksi positif yang benar-benar positif.
<i>Recall</i>	Mengukur kemampuan model dalam menemukan semua data yang benar-benar relevan (positif) dari seluruh data positif yang tersedia.
<i>F1-score</i>	Mencari keseimbangan antara <i>precision</i> dan <i>recall</i> .
<i>ROC Curve</i>	Mengevaluasi kinerja model klasifikasi, terutama dalam kasus klasifikasi biner serta menunjukkan <i>trade-off</i> antara <i>True Positive Rate (TPR)</i> dan <i>False Positive Rate (FPR)</i> pada berbagai ambang batas (<i>Threshold</i>).
<i>Cross Validation</i>	Mengevaluasi performa model secara akurat dan menghindari <i>overfitting</i> , dengan cara membagi data latih menjadi beberapa bagian (<i>fold</i>) dan mengujinya secara bergantian.
<i>Best Hyperparameter</i>	Kombinasi nilai-nilai hiperparameter dalam model <i>machine learning</i> yang menghasilkan kinerja optimal berdasarkan proses evaluasi dari <i>cross validation</i> .

INTISARI

Penelitian ini dilandasi oleh meningkatnya kasus penyakit paru-paru yang erat kaitannya dengan gaya hidup yang kurang sehat serta adanya faktor keturunan. Dalam upaya untuk mengurangi risiko dan mendorong deteksi dini, studi ini bertujuan untuk mengembangkan sistem prediksi penyakit paru-paru berdasarkan data gaya hidup dan penyakit bawaan, dengan memanfaatkan algoritma *Logistic regression*.

Proses penelitian mencakup tahapan pengumpulan data, praproses data, pelatihan model, hingga evaluasi kinerja model prediksi. Dataset yang digunakan meliputi berbagai variabel gaya hidup seperti kebiasaan merokok, aktivitas fisik, pola makan, dan riwayat penyakit dalam keluarga. Algoritma *Logistic regression* dipilih karena mampu menangani klasifikasi dua kelas serta memberikan pemahaman mengenai pengaruh masing-masing variabel terhadap prediksi akhir.

Berdasarkan hasil yang diperoleh, model *Logistic regression* berhasil mencapai tingkat akurasi sebesar 89%, dengan nilai *precision*, *recall*, dan *f1-score* yang juga menunjukkan performa yang baik. Temuan ini menunjukkan bahwa model dapat diandalkan sebagai alat bantu dalam mendeteksi penyakit paru-paru sejak dini, dengan mengandalkan data gaya hidup dan penyakit bawaan.

Dengan demikian, sistem prediksi yang dibangun melalui algoritma ini dinilai efektif dan efisien untuk mendukung pengambilan keputusan medis secara dini. Diharapkan pengembangan lanjutan dapat mendorong pemanfaatan teknologi *data science* dalam pelayanan kesehatan preventif.

Kata kunci: Prediksi, Regresi Logistik, Penyakit paru-paru, gaya hidup, Penyakit bawaan.

ABSTRACT

This study is based on the increasing cases of lung disease which are closely related to unhealthy lifestyles and hereditary factors. In an effort to reduce the risk and encourage early detection, this study aims to develop a lung disease prediction system based on lifestyle data and congenital diseases, by utilizing the Logistic regression algorithm.

The research process includes the stages of data collection, data preprocessing, model training, and evaluation of the performance of the prediction model. The dataset used includes various lifestyle variables such as smoking habits, physical activity, diet, and family history of disease. The Logistic regression algorithm was chosen because it is able to handle two-class classification and provides an understanding of the influence of each variable on the final prediction.

Based on the results obtained, the Logistic regression model managed to achieve an accuracy level of 89%, with precision, recall, and f1-score values also showing good performance. These findings indicate that the model can be relied on as a tool in detecting lung disease early, by relying on lifestyle data and congenital diseases.

Thus, the prediction system built through this algorithm is considered effective and efficient to support early medical decision-making. It is hoped that further development can encourage the use of data science technology in preventive health services.

Keyword: Prediction, Logistic regression, Lung disease, Lifestyle, Congenital disease.