

BAB I PENDAHULUAN

1.1 Latar Belakang

Kanker merupakan penyakit yang dapat membunuh orang di berbagai negara. 9,6 juta orang di seluruh dunia kehilangan nyawa karena kanker [1]. Paru-paru merupakan sistem pernapasan yang digunakan manusia sangat penting bagi kemampuan tubuh untuk mendapatkan oksigen yang dibutuhkan. Selain itu, paru-paru berfungsi sebagai tempat pertukaran karbon dioksida dari darah dengan oksigen dari udara sekitar. Dalam keadaan tertentu, paru-paru mungkin mengalami gangguan yang mengganggu fungsi sistem pernapasan [2].

Indonesia mengalami peningkatan 34.783 kasus kanker paru-paru, dengan angka kematian 30.843 kasus. Orang mungkin kesulitan untuk mengidentifikasi penyakit paru-paru yang dideritanya karena banyaknya jenis penyakit paru-paru dan gejalanya yang hampir sama [3]. Rokok kini semakin populer di kalangan kalangan muda maupun tua di Indonesia. Sekalipun jumlah penggemar rokok meningkat, mereka yang kurang mampu mulai mengembangkan kecanduan nikotin melalui kontak sosial. Perokok mengabaikan risiko kesehatan yang terkait dengan merokok karena dampaknya yang tidak langsung terasa [2].

Karena tingginya kasus penyakit paru-paru, maka untuk dapat mengatasi permasalahan tersebut telah banyak dilakukan penelitian dalam bidang ilmu komputer [4]. *Logistic regression* merupakan pendekatan pembelajaran mesin yang didasarkan pada teori probabilitas ini sering digunakan untuk mengatasi masalah kategorisasi [5]. Jika dibandingkan dengan metode lainnya, regresi logistik memberikan akurasi yang lebih baik pada interpretasi dan pemahaman hubungan antar variabel (gaya hidup dan penyakit bawaan) dengan penyakit paru-paru.

Tujuan dari kajian ini adalah untuk menganalisis pengaruh gaya hidup seperti usia, jenis kelamin, merokok, aktivitas begadang, aktivitas olahraga, penyakit bawaan dan hasil terhadap risiko penyakit paru-paru, serta mengembangkan model prediktif menggunakan algoritma *logistic regression*.

Penelitian ini diharapkan dapat meningkatkan kesadaran masyarakat akan pentingnya gaya hidup sehat dan mencegah faktor penyakit bawaan. Penelitian ini juga mendorong terciptanya solusi kesehatan yang didukung oleh kecerdasan buatan yang membantu para peneliti, profesional medis, dan sektor asuransi untuk menciptakan rencana kesehatan yang lebih efisien.

1.2 Rumusan Masalah

1. Apa saja variabel yang paling signifikan dalam memprediksi penyakit paru-paru menggunakan algoritma *logistic regression*?
2. Seberapa efektif algoritma *logistic regression* dalam memprediksi risiko penyakit paru-paru berdasarkan variabel gaya hidup dan penyakit bawaan?
3. Bagaimana pengaruh penyakit bawaan terhadap kemungkinan seseorang menderita penyakit paru-paru?

1.3 Batasan Masalah

Agar judul penelitian tidak ditafsirkan terlalu luas, berikut ini adalah batasan penelitian:

1. Penelitian ini menggunakan variabel yang relevan seperti:
 - a) Usia berperan karena seiring bertambahnya umur, risiko penurunan fungsi paru-paru meningkat secara alami atau akibat kebiasaan hidup yang kurang sehat.
 - b) Perbedaan jenis kelamin turut memengaruhi risiko, di mana pria umumnya lebih banyak terpapar faktor seperti merokok, sementara secara biologis, wanita bisa jadi lebih rentan terhadap efek dari polusi dan asap rokok.
 - c) Aktivitas merokok, baik secara langsung maupun melalui paparan asap rokok orang lain, menjadi faktor risiko utama untuk penyakit paru-paru seperti PPOK dan kanker paru, serta termasuk variabel paling berpengaruh dalam prediksi.
 - d) Kebiasaan begadang atau kurang tidur dalam jangka panjang dapat melemahkan sistem kekebalan tubuh dan memperparah kondisi

penyakit kronis, termasuk yang berkaitan dengan paru-paru. Meskipun dampaknya tidak bersifat langsung, faktor ini tetap memiliki relevansi dalam konteks klinis.

- e) Olahraga teratur terbukti mampu memperkuat daya tahan tubuh dan memperbaiki fungsi paru-paru. Karena itu, individu yang sering berolahraga cenderung memiliki sistem pemapasan yang lebih sehat dibandingkan yang tidak.
 - f) Kondisi medis yang bersifat bawaan, seperti gangguan pernapasan atau jantung, berpotensi mempertinggi risiko munculnya penyakit paru-paru. Oleh sebab itu, fitur ini sangat relevan dan perlu dicantumkan dalam model analisis prediktif.
2. Menghapus variabel yang tidak relevan seperti:
- a) Variabel pekerjaan tidak menunjukkan hubungan langsung dengan risiko penyakit paru-paru. Meski demikian, jika konteks lingkungan kerja seperti area industri atau tambang disertakan, fitur ini berpotensi signifikan. Namun, data tersebut tidak tercantum dalam dataset.
 - b) Menjadi ibu rumah tangga tidak berkorelasi langsung dengan risiko penyakit paru-paru. Meski potensi pengaruh bisa muncul dari paparan asap dapur atau pencemaran udara di lingkungan rumah, informasi tersebut tidak terwakili dalam dataset.
 - c) Kepemilikan asuransi mencerminkan akses terhadap layanan medis dan kondisi sosial ekonomi, namun keterbatasan data mengenai jenis dan manfaat asuransi menghambat interpretasi yang akurat dalam prediksi penyakit paru-paru.

Pemilihan 7 variabel dalam penelitian ini yaitu usia, jenis kelamin, merokok, aktivitas begadang, aktivitas olahraga, penyakit bawaan, dan hasil didasarkan pada pertimbangan medis serta referensi dari studi sebelumnya yang menunjukkan bahwa faktor-faktor tersebut memiliki keterkaitan kuat dengan kesehatan paru-paru. Penyederhanaan dari 11 atribut menjadi 7

dilakukan melalui proses *feature selection* untuk mengurangi gangguan data (noise), menyederhanakan kompleksitas model, serta meningkatkan ketepatan dan efisiensi dalam proses prediksi.

3. Ruang Lingkup Data

- a) Penelitian ini menggunakan dataset *predic* terkena penyakit paru paru yang diambil dari Kaggle.
- b) Data *predic* yang tidak lengkap (*missing values*) atau tidak relevan akan dihapus atau diolah untuk menjaga kualitas analisis.

4. Metode yang digunakan

- a) Penelitian ini menggunakan Algoritma *Logistic regression* untuk memprediksi penyakit berdasarkan usia, jenis kelamin, merokok, aktivitas begadang, aktivitas olahraga, penyakit bawaan dan hasil karena cocok untuk tipe masalah klasifikasi biner, memberikan interpretasi yang baik, tidak mengharuskan asumsi linearitas, dapat mengelola variabel kategori, mencegah *overfitting* dengan regularisasi, memiliki performa yang baik pada data kategori, dan mampu menangani ketergantungan variabel, sesuai dengan karakteristik, asumsi dalam analisis kesehatan dan prediksi penyakit.
- b) Metode yang digunakan menggunakan Algoritma *Logistic regression* tanpa membandingkan performa dari Algoritma lainnya.

1.4 Tujuan Penelitian

1. Menganalisis dan mengidentifikasi variabel-variabel yang memiliki pengaruh paling besar terhadap prediksi penyakit paru-paru dalam model yang dibangun.
2. Mengidentifikasi efektivitas algoritma *logistic regression* dalam memprediksi penyakit paru-paru berdasarkan gaya hidup dan penyakit bawaan.
3. Menentukan sejauh mana pengaruh penyakit bawaan terhadap kemungkinan seseorang terkena penyakit paru-paru.

1.5 Manfaat Penelitian

Manfaat dari penelitian ini adalah meningkatkan kesadaran masyarakat akan pentingnya menjalani gaya hidup sehat, memberikan kontribusi ilmiah sebagai panduan untuk penelitian dimasa depan dan memberikan informasi yang berguna bagi para profesional medis saat memberikan nasihat kepada pasien. Selain itu, temuan penelitian ini dapat membantu pembuat kebijakan membuat program untuk mencegah penyakit paru-paru dan membuka jalan untuk penyelidikan di masa depan terhadap variabel tambahan terkait kesehatan paru-paru dan penciptaan teknis analisis yang lebih canggih.

1.6 Sistematika Penulisan

Laporan skripsi ini terdiri dari 5 bab, dengan sistematika penulisan sebagai berikut:

BAB I PENDAHULUAN

Bab ini menjelaskan latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, serta sistematika penulisan yang digunakan dalam skripsi ini.

BAB II TINJAUAN PUSTAKA

Bab ini menjelaskan studi literatur dan dasar teori yang digunakan meliputi *data mining*, *logistic regression*, *preprocessing*, *data splitting*, *prediction*, *confusion matrix*, *accuracy*, *precision*, *recall*, *f1-score*, *cross validation*, *hyperparameter* dan *python*.

BAB III METODE PENELITIAN

Bab ini menjelaskan objek penelitian, alur penelitian, alat dan bahan.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menjelaskan hasil dan pembahasan.

BAB V PENUTUP

Bab ini berisikan kesimpulan dan saran.