|  |  |
| --- | --- |
|  |  |
| **LAPORAN TUGAS COMPUTATIONAL BIOLOGY**    **PREDIKSI KANKER PARU-PARU MENGGUNAKAN MODEL MACHINE LEARNING**   |  |  | | --- | --- | | **Stanley Raditya** | **2602103671** | | **Jenibelle Wievin** | **2602105632** | | **Mokhammad Mulkie Al-Ghani** | **2602128781** | |  |  |   **BINUS UNIVERSITY**  **2024** | |

DAFTAR ISI

Halaman Judul Penelitian

Daftar Isi

Daftar Gambar

Daftar Tabel

Bab 1. Pendahuluan

Bab 2. Tinjauan Pustaka

Bab 3. Metode Penelitian

Bab 4. Hasil dan Pembahasan

Bab 5. Kesimpulan dan Saran

Daftar Pustaka

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1. PENDAHULUAN

1. **LATAR BELAKANG**

Kanker merupakan masalah yang terlihat dan terkesan sangat kompoleks. Tetapi pada intinya, kanker merupakan sebuah permasalahan yang sebenarnya sederhana. Kanker adalah kondisi dimana sel-sel dalam tubuh kita bertumbuh dan bertambah banyak tak terkendali disaat tubuh kita tidak membutuhkannya, dimana pada kondisi seharusnya sel normal bertumbuh dan mati disaat tubuh kita berkehendak demikian saja.

Tahap selanjutnya kanker akan membentuk gumpalan jaringan yang menumpuk atau benjolan yang umumnya disebut sebagai ‘tumor’. Sel kanker merupakan sel tubuh yang abnormal sehingga tidak bisa dikendalikan oleh tubuh kita dan tumor tadi dapat mengahalangi tubuh kita untuk bekerja dengan baik dan seiring berjalannya waktu, tumor ini dapat menyebar ke bagian-bagian tubuh lain yang menempel. Tumor juga tergolong menjadi 2, tumor ganas dan tumor jinak. Tumor jinak tidak akan menyebar ke jaringan di dekatnya. Saat dokter mengangkatnya, massa yang tidak berbahaya tersebut biasanya tidak tumbuh kembali sedangkan tumor ganas akan tumbuh kembali dan bisa menyerang anggota tubuh lainnya. Dalam beberapa kasus tumor jinak juga bisa mengambil nyawa orang contohnya tumor jinak di otak.

Kanker bisa disbabkan oleh faktor internal dan juga faktor eksternal. Faktor risiko internal trejadi jika seseorang memiliki keturunan atau riwayat penderita kanker dalam keluarganya. Sedangkan faktor eksternal ada beberapa seperti perubahan hormon dalam tubuh, sering terpapar bahan kimia berbahaya (asbes dan benzena), pola makan dan hidup seperti konsumsi alkohol dalam jumlah berlebihan, merokok, obesitas, sinar matahari berlebihan, dan masih banyak lagi.

Sel kanker bisa tumbuh dimanapun, dan kali ini kita kan memfokuskan kepada paru-paru. Kanker paru-paru merupakan suatu kondisi disaat sel-sel tumbuh tidak terkendali di dalam organ paru-paru.

Kanker paru-paru merupakan salah satu penyakit yang penyakit mematikan di dunia, dengan tingkat kematian yang terbilang cukup tinggi dan jumlah kasus yang banyak di banyak negara. Meskipun telah ada upaya besar dalam pencegahan dan pengobatannya, deteksi dini masih menjadi kunci dalam pencegahan dan tingkat keberlangsungan hidup pasien atau pengidap.

Kami mengangkat tema ini juga dikarenakan pengamatan kalau semakin kebelakang semakin banyak anak-anak muda yang merokok dan memulai merokok dari usia yang bisa terbilang dini atau masih dibawah usia 18 tahun, dibandingkan oleh generasi-generasi sebelum. Dimulai dari anak-anak umur 5-7 tahun yang sudah mulai merokok, ini juga dikarenakan oleh kurangnya perhatian dari orang tua juga. Diluar itu juga lingkungan sekitar saya juga banyak yang merokok secara tradisional maupun elektrik.

Dari kondisi sekitar tersebutlah saya mulai berpikir akan risiko yang mungkin saja bisa terjadi. Tidak dalam waktu dekat mungkin tapi kita tidak tahu dimasa depan nantinya. Karena reaksi tubuh dengan senyawa-senyawa kimia dan imun tubuh setiap orang juga berbeda-beda.

1. **RUMUSAN MASALAH**

Dari latar belakang tersebut kami merumuskan beberapa masalah sebagai berikut:

* Proses pembuatan model *machine learning* untuk memprediksi kemungkinan kanker paru-paru berdasarkan beberapa variabel yang ada seperti alkohol, merokok, alergi, dll.
* Hasil dan evaluasi dari model *machine learning* dalam memprediksi seseorang mengidap kanker paru-paru berdasarkan faktor-faktor yang diberikan.
* Tingkat akurasi model *machine learning* dengan membandingkan dengan hasil penelitian lain secara medis

1. **TUJUAN MASALAH**

* Mengidentifikasi faktor-faktor apa saja yang kemungkinan besar dapat menyebabkan kanker paru-paru dengan menggunakan *model machine learning*.
* Mencari model yang paling cocok dengan permasalahan ini, serta mengevaluasi hasil serta membuat analisa dari performa model *machine learning* yang dibuat
* Kami berharap dengan penelitian ini orang-orang yang terlibat aktivitas dan faktor-faktor yang dapat menyebabkan kanker paru-paru seperti merokok bisa lebih tersadar dan minimal bisa membuatnya mengurangi merokoknya.

1. **MANFAAT PENELITIAN**
2. Bagi peneliti

* Lebih memahami dan memperbanyak pengalaman dan latihan menggunakan *machine learning* untuk penerapannya dalam kasus di lapangan.

1. Bagi masyarakat

* Penelitian ini juga dapat meninggkatkan kesadaran masyarakat tentang pentingnya deteksi dini terhadap kanker paru-paru dan pemahaman akan risikonya, sehingga diharapkan adanya peningkatan kesadaran untuk pencegahan dan lebih memperhatikan kesehatan tubuh sendiri dan

BAB 2. TINJAUAN PUSTAKA

1. ***MACHINE LEARNING***

*Machine Learning* atau pembelajaran mesin merupakan penerapan pada kecerdasan buatan yang memebrikan sistem kemampuan belajar secara otomatis dari sekumpulan data untuk melakukan tugas terntentu tanpa di program secara eksplisit (Batta, 2020). Sedangkan menurut (Rebala dkk., 2019), *machine learning* merupakan bidang ilmu computer yang mempelajari algoritma dan Teknik untuk menghasilkan sebuah solusi pada masalah kompleksitas yang sulit dilakukan dengan menggunakan metode pemrograman konvensional. Pada prosesna algoritma pada *machine learning* akan mencari pola tertentu pada setiap data atau yyang disebut dengan *dataset* untuk menghasilkan sebuah *rules*. Secara umum metode pada *machine learning* dibagi menjadi empat tipe berdasarkan cara pembelajarannya (Yu & He, 2019), yaitu:

* 1. *Supervised Learning*

Supervised learning merupakan metode pada machine learning yang memiliki fungsi dalam memetakan data masukan dan keluaran berdasarkan contoh yang tersedia atau yang disebut dataset. Dataset masukan terdiri dari data latih dan data uji, dimana data latih memiliki variabel keluaran yang perlu diprediksi atau diklasifikasi. Pada prosesnya metode ini akan mempelajari pola dari dataset latih dan menerapkannya ke setiap dataset uji untuk melakukan prediksi atau klasifikasi.

* 1. *Unsupervised Learning*

Unsupervised learning merupakan metode pada machine learning yang digunakan untuk melakukan pengelompokan data tanpa pengawasan, artinya dataset yang dikelola tidak mempunyai label tertentu. Pada prosesnya metode ini akan mengeksplorasi dan menemukan struktur dalam data untuk mengenali kelas data. Teknik yang digunakan pada computer vision yaitu density estimation (estimasi kepadatan) dan dimensionality reduction (reduksi dimensi) yang mengarah ke berbagai aspek pada visual features (fitur visual) (Caron dkk., 2019).

* 1. *Semi – supervised Learning*

Semi-supervised learning merupakan kombinasi metode supervised learning dan unsupervised learning, artinya data yang memiliki label dan tanpa label akan digunakan untuk melatih algoritma agar dapat melakukan proses pembelajaran sendiri. Metode ini digunakan untuk data yang memiliki label dengan jumlah yang sedikit dan data tanpa label dalam jumlah yang lebih banyak.

* 1. *Reinforcement Learning*

Reinforcement learning merupakan metode pada machine learning yang berhubungan dengan pengambilan keputusan secara berurutan untuk memperolah hasil yang terbaik. Pada prosesnya metode ini akan mempelajari struktur data yang dapat memodifikasi atau memperoleh perilaku dan keterampilan baru secara bertahap untuk mendapatkan reward berdasarkan trial and error experience (pengalaman coba- coba) (François-lavet dkk., 2018).

1. ***DEEP LEARNING***

*Deep learning* atau pembelajaran mendalam merupakan sub-bidang yang spesifik dari machine learning dengan penerapan algoritma artificial neural network (*neural network* tiruan) untuk mempelajari respresentasi data yang tidak terstruktur dalam jumlah yang besar (Pramestya, 2018). Pada *deep learning* lapisan representasi dipelajari melalui model yang disebut neural network (*neural network*) dalam lapisan literal yang ditumpuk satu sama lain secara terstruktur. Terdapat tiga lapisan utama pada *deep learning*, diantaranya input layer, hidden layer dan output layer. Berikut merupakan ilustrasi lapisan arsitektur *neural network* pada *deep learning*.



Arsitektur *deep learning* diatas disebut sebagai multi layer perceptron, yang memiliki 4 node neuron pada input layer (lapisan masukan) berfungsi sebagai jalur untuk mentransmisikan sinyal masukan ke node berikutnya, 2 hidden layer (lapisan tersembunyi) yang merupakan perantara antara input layer dengan output layer dan berfungsi untuk menambah instruksi logika, serta 3 node II-4 neuron pada output layer (lapisan keluaran) yang menjadi hasil akhir dari neural network. Jumlah neuron yang diterapkan pada sebuah lapisan tersembunyi sangat mempengaruhi performa *deep learning*. Jika jumlah node neuron yang diterapkan pada arsitektur *deep learning* terlalu sedikit maka dapat menyebabkan model menjadi underfitting. Jika jumlah node neuron yang diterapkan terlalu banyak maka dapat menyebabkan model menjadi overfitting. Penelitian ini menerapkan teknik *deep learning* pada metode machine learning yaitu supervised learning untuk melakukan klasifikasi HAR pada video, dengan menggunakan kombinasi algoritma CNN dan LSTM.

BAB 3. METODE PENELITIAN

Penelitian ini bertujuan untuk menggunakan model machine learning dalam memprediksi kemungkinan seseorang mengidap kanker paru-paru berdasarkan sejumlah variabel yang relevan. Metode penelitian ini dibagi menjadi beberapa tahap yang mencakup: Pengumpulan Data, Pra-pemrosesan Data, Pemilihan dan Pelatihan Model Machine Learning, Evaluasi Model, dan Analisis Hasil.

**1. Pengumpulan Data**

Langkah pertama dalam penelitian ini adalah mengumpulkan dataset yang relevan yang akan digunakan untuk pelatihan dan pengujian model machine learning. Sumber data yang potensial meliputi:

*- Basis data medis public :*

Misalnya, UCI Machine Learning Repository yang menyediakan data medis untuk keperluan penelitian.

*- Rumah sakit atau klinik :*

Data riwayat medis pasien dari fasilitas kesehatan yang memiliki catatan yang rinci.

*- Survei atau studi sebelumnya :*

Penelitian lain yang telah dilakukan dan memiliki data yang dapat digunakan.

Data yang dikumpulkan harus mencakup berbagai variabel penting, seperti:

*- Kebiasaan merokok :*

Informasi mengenai jumlah rokok yang dikonsumsi per hari, durasi merokok dalam tahun, dll.

*- Konsumsi alcohol :*

Tingkat konsumsi alkohol dan frekuensinya.

*- Riwayat Kesehatan :*

Keluarga yang mengidap kanker, riwayat penyakit paru-paru lainnya, dll.

*- Faktor lingkungan :*

Paparan terhadap bahan kimia berbahaya, tingkat polusi udara, dll.

*- Data demografi :*

Informasi seperti usia, jenis kelamin, status pekerjaan, dll.

*- Gejala awal :*

Gejala-gejala yang dialami seperti batuk kronis, sesak napas, nyeri dada, dll.

**2. Pra-pemrosesan Data**

Tahap ini sangat penting untuk memastikan bahwa data yang digunakan bersih, konsisten, dan siap untuk dianalisis. Langkah-langkah yang diambil meliputi:

*- Pembersihan Data :*

Mengidentifikasi dan menangani data yang hilang (missing values), menghapus data duplikat, serta memperbaiki data yang tidak konsisten atau tidak relevan.

*- Normalisasi dan Standarisasi :*

Mengubah skala variabel sehingga berada dalam rentang yang sama untuk meningkatkan performa model, seperti menggunakan teknik min-max scaling atau z-score normalization.

*- Feature Engineering :*

Membuat fitur-fitur baru yang relevan dari data yang ada, misalnya menghitung total paparan rokok dalam satu tahun atau indeks risiko berdasarkan kombinasi beberapa faktor.

*- Pembagian Data :*

Membagi dataset menjadi set pelatihan (training set) dan set pengujian (test set) dengan proporsi yang sesuai, misalnya 80% data untuk pelatihan dan 20% untuk pengujian, untuk memastikan model dapat diuji dengan data yang belum pernah dilihat sebelumnya.

**3. Pemilihan dan Pelatihan Model Machine Learning**

Pada tahap ini, beberapa algoritma machine learning akan dipilih dan dilatih untuk memprediksi kanker paru-paru. Algoritma yang akan dieksplorasi meliputi:

*- Regresi Logistik :*

Digunakan untuk mengukur probabilitas biner dari kejadian kanker paru-paru.

*- Decision Trees :*

Membantu memahami hubungan antara variabel input dan hasil prediksi dengan struktur pohon keputusan yang mudah diinterpretasikan.

*- Random Forest :*

Kombinasi dari beberapa pohon keputusan untuk meningkatkan akurasi prediksi dan mengurangi risiko overfitting.

*- Support Vector Machines (SVM) :*

Digunakan untuk klasifikasi data yang kompleks dengan margin maksimal yang memisahkan kelas-kelas yang berbeda.

*- Neural Networks :*

Khususnya *deep learning*, yang mampu memodelkan hubungan non-linear yang kompleks antara variabel input dan output.

Proses pelatihan model melibatkan beberapa langkah berupa :

*- Melatih model :*

Menggunakan data pelatihan untuk melatih setiap algoritma yang dipilih.

*- Hyperparameter tuning :*

Menggunakan teknik seperti Grid Search atau Random Search untuk menemukan kombinasi hyperparameter terbaik yang dapat meningkatkan performa model.

*- Cross-validation :*

Menggunakan teknik seperti K-Fold Cross-Validation untuk memastikan model tidak overfitting dan performa yang dihasilkan stabil pada berbagai subset data.

**4. Evaluasi Model**

Evaluasi model dilakukan untuk mengukur seberapa baik model yang telah dilatih dapat memprediksi kanker paru-paru. Beberapa metode evaluasi yang digunakan meliputi:

*- Confusion Matrix :*

Matriks yang digunakan untuk menghitung jumlah True Positive, True Negative, False Positive, dan False Negative, sehingga performa model dapat dianalisis lebih mendetail.

*- Accuracy, Precision, Recall, dan F1-Score :*

Metode-metode ini digunakan untuk mengukur kinerja model secara keseluruhan, di mana akurasi menunjukkan seberapa sering model membuat prediksi yang benar, presisi menunjukkan proporsi prediksi positif yang benar, recall menunjukkan proporsi kasus positif yang berhasil diidentifikasi oleh model, dan F1-Score adalah rata-rata harmonis dari presisi dan recall.

*- ROC Curve dan AUC (Area Under Curve) :*

Mengukur kemampuan model dalam membedakan antara kelas positif dan negatif dengan memplot kurva Receiver Operating Characteristic (ROC) dan menghitung area di bawah kurva tersebut.

*- Cross-validation Score :*

Untuk memastikan model bekerja dengan baik pada data yang belum pernah dilihat sebelumnya dan memberikan estimasi performa model yang lebih reliabel.

**5. Analisis Hasil**

Setelah model dievaluasi, langkah berikutnya adalah menganalisis hasil untuk memahami faktor-faktor yang berpengaruh signifikan terhadap prediksi kanker paru-paru. Analisis ini meliputi:

*- Feature Importance :*

Mengidentifikasi fitur mana yang memiliki pengaruh paling besar terhadap prediksi model, yang dapat memberikan wawasan tentang faktor risiko utama kanker paru-paru.

*- Error Analysis :*

Mengidentifikasi dan menganalisis kesalahan prediksi untuk memahami mengapa model gagal dalam kasus-kasus tertentu dan bagaimana perbaikan dapat dilakukan.

*- Comparative Analysis :*

Membandingkan hasil penelitian ini dengan penelitian atau model medis lainnya untuk memastikan validitas dan keandalan model yang dikembangkan.

**6. Implementasi dan Interpretasi**

Setelah model berhasil dilatih dan dievaluasi, langkah selanjutnya adalah implementasi dalam skenario nyata. Implementasi ini bertujuan untuk:

*- Penggunaan oleh profesional medis :*

Model dapat digunakan oleh dokter dan ahli kesehatan untuk membantu dalam deteksi dini kanker paru-paru dan perencanaan pengobatan yang lebih efektif.

*- Meningkatkan kesadaran Masyarakat :*

Memberikan rekomendasi kepada masyarakat tentang faktor risiko kanker paru-paru dan pentingnya pencegahan melalui gaya hidup sehat.

*- Mendorong penelitian lanjutan :*

Hasil penelitian ini diharapkan dapat menjadi dasar bagi penelitian lanjutan untuk mengembangkan model yang lebih baik dan lebih akurat dalam memprediksi kanker paru-paru.

Dengan metode penelitian yang komprehensif ini, diharapkan prediksi kanker paru-paru menggunakan model machine learning dapat menjadi lebih akurat, sehingga dapat membantu dalam deteksi dini dan pencegahan penyakit tersebut, serta memberikan kontribusi signifikan dalam bidang kesehatan masyarakat.

BAB 4. HASIL EKSPERIMEN

A screen shot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

A black screen with a black border

Description automatically generated

A screen shot of a computer program

Description automatically generated

A graph of a number of bars

Description automatically generated with medium confidence

A graph with blue and orange squares

Description automatically generatedA screenshot of a graph

Description automatically generated

A graph with blue and orange bars

Description automatically generated

A graph with blue and orange bars

Description automatically generated

A graph with blue and orange bars

Description automatically generated

A graph of cancer

Description automatically generated

A graph of different colored squares

Description automatically generatedA graph with blue and orange bars

Description automatically generated

A graph of cancer

Description automatically generated with medium confidenceA graph of a cancer

Description automatically generated with medium confidence

A graph of cancer

Description automatically generated

A graph with blue and orange squares

Description automatically generated

A graph of cancer

Description automatically generated

A graph of cancer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a chart

Description automatically generated

Dari heatmap yang ada diatas ini kitab isa melihat korelasi yang ada antar semua variable pada dataset ini. Bisa kita lihat sendiri kalau smoking bahkan lebih rendah daripada umur sebagai dalam segi korelasinya terhadap kanker paru-paru. Bisa dibilang kalau merokok belum tentu menjadi penyebab utama dari kanker paru-paru. Hal – hal yang tidak terlalu diduga-duga menjadi salah satu penyebab besar akan kanker paru-paru yaitu Alergi dengan korelasi terbesar diangkan 0.33 diikuti dengan konsumsi alkohol diurutan kedua.

Hasil pendekatan model machine learning menggunakan Logistic Regression:

A screen shot of a graph

Description automatically generated

Hasil model machine learning dengan pendekatan SVM

A screenshot of a computer

Description automatically generated

BAB 5. KESIMPULAN

Kesimpulan eksperimen yang dilakukan:

1. **Logistic Regression (ROC AUC Score: 0.95)**: Model logistic regression memiliki kinerja yang baik dengan skor ROC AUC sebesar 0.95. ROC AUC adalah metrik evaluasi yang umum digunakan untuk mengukur kualitas prediksi model klasifikasi, di mana nilai mendekati 1 menunjukkan kinerja yang sangat baik.
2. **SVM (ROC AUC Score: 0.97)**: Model Support Vector Machine (SVM) menunjukkan kinerja yang lebih tinggi dibandingkan dengan logistic regression, dengan skor ROC AUC sebesar 0.97. Hal ini menandakan bahwa SVM memiliki kemampuan yang lebih baik dalam memisahkan kelas yang berbeda dalam dataset.

Penjelasan dari heatmap menunjukkan bahwa terdapat beberapa temuan menarik terkait korelasi antar variabel dalam dataset:

* **Smoking bukanlah penyebab utama kanker paru-paru**: Meskipun merokok seringkali diasosiasikan dengan kanker paru-paru, heatmap menunjukkan bahwa korelasi antara merokok dan kanker paru-paru lebih rendah dibandingkan dengan faktor lainnya seperti umur. Hal ini menunjukkan bahwa meskipun merokok berperan dalam risiko kanker paru-paru, faktor-faktor lain juga berpengaruh secara signifikan.
* **Alergi memiliki korelasi yang signifikan dengan kanker paru-paru**: Salah satu temuan mengejutkan dari heatmap adalah bahwa alergi memiliki korelasi yang tinggi (0.33), menempatkannya sebagai faktor risiko yang signifikan untuk kanker paru-paru. Hal ini menyoroti pentingnya mempertimbangkan faktor-faktor lain selain merokok dalam pencegahan dan deteksi dini kanker paru-paru.
* **Konsumsi alkohol juga berperan**: Selain alergi, konsumsi alkohol juga memiliki korelasi yang cukup tinggi dengan kanker paru-paru, meskipun di urutan kedua setelah alergi. Ini menunjukkan bahwa gaya hidup dan faktor lingkungan lainnya juga memainkan peran penting dalam risiko terhadap kanker paru-paru.

Dengan demikian, kesimpulan utama adalah bahwa meskipun merokok memiliki korelasi dengan kanker paru-paru, faktor-faktor lain seperti alergi dan konsumsi alkohol juga memiliki pengaruh yang signifikan. Oleh karena itu, pendekatan dalam pencegahan dan deteksi kanker paru-paru harus holistik dan mempertimbangkan berbagai faktor risiko yang ada.