Evaluasi Model Multiple Machine Learning dan Metode Sampling Data untuk Prediksi Infeksi Entamoeba di Indonesia

Pendahuluan

Entamoeba histolytica (E. histolytica) merupakan agen patogen utama penyebab penyakit amebiasis, yang menjadi isu penting dalam kesehatan masyarakat global. Penyakit ini menyebabkan tingkat kesakitan dan kematian yang tinggi, terutama di negara-negara berkembang. Selain E. histolytica, parasit protozoa intestinal lainnya, E. dispar dan E. moshkovskii, juga perlu diwaspadai meskipun mereka merupakan amoeba non-patogen. Meskipun mereka tidak menyebabkan penyakit, keberadaan mereka dapat menunjukkan adanya kontaminasi feses dalam sumber air atau makanan, yang bisa menjadi indikator risiko infeksi E. histolytica.

Indonesia, sebagai negara kepulauan, memiliki banyak pulau besar dan kecil. Banyak dari pulau-pulau ini berada di bagian terluar Indonesia dan sebagian besar adalah pulau-pulau kecil dan sebagian diantaranya adalah pulau berpenduduk. Pulau Weh, sebagai contoh, berbatasan langsung dengan perairan India, Malaysia, dan Thailand. Kondisi geografis ini, ditambah dengan kurangnya infrastruktur sanitasi yang memadai, dapat meningkatkan risiko penyebaran infeksi E. histolytica.

Namun, informasi tentang prevalensi dan faktor-faktor yang mempengaruhi infeksi E. histolytica masih kurang, sehingga belum ditemukannya kebijakan khusus terhadap pencegahan dan penanggulangan penyakit yang disebabkan oleh infeksi parasit intestinal ini.

Dalam penelitian ini, kami bertujuan untuk mengestimasi prevalensi, menganalisis faktor risiko, dan mengeksplorasi model-model dalam Multiple Machine Learning. Machine learning adalah metode komputasi yang memungkinkan sistem untuk belajar dari data dan membuat prediksi atau keputusan tanpa perlu diprogram secara eksplisit. Kami memilih untuk mengevaluasi berbagai model machine learning karena kami percaya bahwa pendekatan ini akan memberikan hasil yang lebih baik dalam memprediksi infeksi E. histolytica.

Kami akan menggunakan empat model machine learning dan empat metode sampling data, dan dataset akan dibagi menjadi 10% data uji dan 90% data training. Kami akan membandingkan performa model-model ini menggunakan metrik seperti AUROC, AUPRC, F1 score, akurasi, CV std, dan CV avg untuk menentukan kombinasi model dan metode sampling yang paling efektif. Selain itu, setiap model machine learning juga akan dilatih tanpa menggunakan metode sampling data untuk memahami performa dasar mereka. Dengan demikian, penelitian ini diharapkan dapat memberikan wawasan baru tentang cara terbaik untuk memprediksi infeksi E. histolytica, yang pada akhirnya dapat membantu dalam pengembangan strategi pencegahan dan penanggulangan penyakit amebiasis.

Metode

Jenis dan Sampel Penelitian:

Penelitian ini adalah studi epidemiologi yang menggunakan algoritma Machine Learning. Sampel penelitian adalah masyarakat Pulau Weh yang berumur ≥ 10 tahun. Kami memilih untuk menggunakan metode non-probability sampling karena keterbatasan waktu dan sumber daya. Meskipun metode ini mungkin tidak memberikan representasi yang sempurna dari populasi target, kami berusaha memastikan bahwa sampel kami cukup beragam dan mencakup berbagai demografi dan kondisi kesehatan. Total ada 335 responden yang dipilih berdasarkan kriteria inklusi dan eksklusi tertentu, yang akan dijelaskan lebih lanjut dalam bagian berikutnya.

Pemeriksaan dan Pengukuran:

Pengambilan sampel tinja dilakukan dengan menggunakan metode yang telah ditetapkan oleh WHO. Identifikasi kompleks Entamoeba histolytica/dispar/moshkovskii dilakukan secara mikroskopis dengan menggunakan teknik pewarnaan khusus. Data perilaku kesehatan dan sanitasi lingkungan dikumpulkan melalui wawancara dan observasi langsung. Kami menggunakan kuesioner yang telah divalidasi dan dilatih untuk memastikan keakuratan dan konsistensi data.

Manajemen dan Analisis Data:

Data diolah dan dianalisis menggunakan software statistik dan Machine Learning. Dataset dibagi menjadi 10% data uji dan 90% data pelatihan. Pembagian ini dilakukan untuk memastikan bahwa model Machine Learning kami dapat belajar dari sebagian besar data (data pelatihan) dan kemudian diuji pada data yang belum pernah dilihat sebelumnya (data uji) untuk memastikan keakuratan dan generalisabilitas model.

Penggunaan Machine Learning:

Kami menggunakan 4 model Machine Learning dan 4 metode data sampling. Pilihan model dan metode sampling ini didasarkan pada penelitian sebelumnya dan pertimbangan praktis. Setiap model dilatih baik dengan dan tanpa metode sampling data. Hasil dari training kemudian dianalisis menggunakan metrik seperti AUROC, AUPRC, F1 score, akurasi, CV std, dan CV avg. Analisis ini memungkinkan kami untuk menentukan kombinasi model dan metode sampling yang paling efektif dalam memprediksi infeksi E. histolytica.