Vol. 1, No. 1, (2023) E-ISSN: XXXX-XXXX

Analisis Prediktif Penyakit Menular dengan Menggunakan K-Nearest Neighbors dan Naïve Bayes di Indonesia

Bery Agustianto
Magister Teknik Informatika Universitas Pamulang
e-mail: beryagustianto@gmail.com

Abstrak*—Penyakit menular, khususnya Tuberkulosis (TB), tetap menjadi tantangan kesehatan utama di Indonesia. Penggunaan metode prediktif berbasis pembelajaran mesin dapat membantu dalam mengidentifikasi pola penyebaran penyakit dan mendukung upaya pencegahan yang lebih efektif. Penelitian ini bertujuan untuk menganalisis performa tiga algoritma pembelajaran mesin, yaitu K-Nearest Neighbors (KNN), Decision Tree, dan Naïve Bayes, dalam memprediksi jumlah kasus TB di berbagai provinsi di Indonesia berdasarkan data jumlah penduduk. Data yang digunakan dalam penelitian ini diperoleh dari laporan kesehatan publik dan statistik populasi, mencakup informasi tentang jumlah kasus TB dan jumlah penduduk di berbagai provinsi. Setelah melakukan pra-pemrosesan data, termasuk pembersihan, normalisasi, dan pembagian data menjadi set latih dan uji, ketiga algoritma tersebut diterapkan dan dievaluasi. Hasil penelitian menunjukkan bahwa algoritma Decision Tree memberikan performa terbaik dalam regresi, dengan Mean Absolute Error (MAE) sebesar 11,948.45 dan R-squared (R^2) sebesar 0.6247, menunjukkan kemampuan yang lebih baik dalam menangkap pola dari dataset dibandingkan dengan KNN. Algoritma KNN menunjukkan nilai MAE sebesar 16,298.91 dan R-squared sebesar 0.3424, mengindikasikan performa yang kurang memuaskan. Sementara itu, Naïve Bayes menunjukkan performa yang sangat baik dalam klasifikasi dengan akurasi sempurna (1.0), precision, recall, dan f1-score masing-masing sebesar 1.0. Berdasarkan hasil ini, Decision Tree direkomendasikan untuk prediksi jumlah kasus TB dalam pendekatan regresi, sementara Naïve Bayes sangat cocok untuk tugas klasifikasi. Penelitian ini memberikan kontribusi penting dalam penggunaan algoritma pembelajaran mesin untuk prediksi penyakit menular dan dapat membantu dalam strategi pencegahan dan pengendalian penyakit di Indonesia.

***Kata Kunci*— Tuberkulosis; Prediksi Penyakit; K-Nearest Neighbors; Decision Tree; Naïve Bayes; Pembelajaran Mesin.**

I. PENDAHULUAN

Penyakit menular merupakan salah satu tantangan utama dalam kesehatan masyarakat, baik di tingkat nasional maupun global. Di Indonesia, penyakit menular seperti Tuberkulosis (TB), HIV/AIDS, dan penyakit lainnya masih menjadi masalah kesehatan yang serius. Tuberkulosis, khususnya, adalah salah satu penyakit menular yang paling mematikan di dunia. Menurut laporan dari World Health Organization (WHO), Indonesia termasuk dalam daftar negara dengan beban TB tertinggi, yang mencakup sekitar 8% dari total kasus TB global pada tahun 2020.

Keberhasilan dalam mengendalikan penyebaran penyakit menular sangat bergantung pada kemampuan untuk memprediksi pola penyebarannya dan mengidentifikasi faktor-faktor risiko yang berkontribusi. Dalam beberapa dekade terakhir, perkembangan teknologi informasi dan ilmu data telah membuka peluang baru untuk analisis prediktif dalam bidang kesehatan masyarakat. Algoritma pembelajaran mesin, seperti K-Nearest Neighbors (KNN), Decision Tree, dan Naïve Bayes, menawarkan alat yang kuat untuk memproses data besar dan kompleks serta mengidentifikasi pola yang tidak dapat terlihat secara manual.

Salah satu pendekatan yang efektif dalam analisis prediktif adalah penggunaan data demografis, seperti jumlah penduduk, sebagai variabel penentu dalam model prediksi. Data jumlah penduduk dapat memberikan indikasi yang signifikan tentang potensi penyebaran penyakit menular di berbagai daerah. Dengan memanfaatkan data ini, model prediktif dapat membantu mengidentifikasi wilayah-wilayah dengan risiko tinggi dan memungkinkan intervensi kesehatan yang lebih tepat waktu dan efisien.

Penggunaan algoritma pembelajaran mesin untuk prediksi penyakit menular telah menunjukkan hasil yang menjanjikan dalam berbagai penelitian. Algoritma K-Nearest Neighbors (KNN), misalnya, dikenal karena kesederhanaannya dan kemampuannya dalam menangani data non-linear. Decision Tree, di sisi lain, menawarkan interpretasi yang mudah dan visualisasi yang jelas dari proses pengambilan keputusan . Sementara itu, Naïve Bayes, dengan dasar probabilistiknya, dapat memberikan prediksi yang cepat dan efektif bahkan dengan jumlah data pelatihan yang relatif kecil.

Penelitian ini bertujuan untuk menganalisis prediksi jumlah kasus TB di berbagai provinsi di Indonesia menggunakan algoritma K-Nearest Neighbors (KNN), Decision Tree, dan Naïve Bayes. Dengan memanfaatkan data jumlah penduduk sebagai fitur, penelitian ini berusaha untuk mengembangkan model prediksi menggunakan KNN, Decision Tree, dan Naïve Bayes. Membandingkan performa model berdasarkan metrik evaluasi seperti Mean Absolute Error (MAE) dan R-squared (R^2) untuk regresi, serta akurasi dan Area Under the Curve (AUC) untuk klasifikasi. Memberikan rekomendasi algoritma terbaik untuk prediksi jumlah kasus TB berdasarkan hasil komparasi model. Hasil dari penelitian ini diharapkan dapat memberikan wawasan yang berharga bagi pembuat kebijakan dan praktisi kesehatan dalam merancang strategi pencegahan dan pengendalian penyakit

menular di Indonesia. Selain itu, penelitian ini juga berkontribusi pada literatur ilmiah mengenai penggunaan teknik pembelajaran mesin dalam epidemiologi penyakit menular.

II. METODE PENELITIAN

Penelitian ini menggunakan data sekunder yang diperoleh dari laporan kesehatan publik dan statistik populasi di Indonesia. Dataset ini mencakup informasi tentang jumlah kasus Tuberkulosis (TB) dan jumlah penduduk di berbagai provinsi di Indonesia. Data ini diambil dari laporan tahunan kesehatan dan sensus penduduk yang diterbitkan oleh Badan Pusat Statistik (BPS) dan Kementerian Kesehatan Indonesia.

Data yang diperoleh mengalami beberapa langkah pra-pemrosesan untuk memastikan kualitas dan konsistensi sebelum digunakan dalam analisis prediktif. Langkah-langkah pra-pemrosesan data meliputi:

1. Pembersihan Data: Menghapus atau mengimputasi nilai yang hilang, mengatasi data duplikat, dan memperbaiki inkonsistensi dalam data.
2. Transformasi Data: Normalisasi variabel numerik untuk memastikan bahwa semua fitur berada dalam skala yang sama, yang penting untuk algoritma pembelajaran mesin.
3. Pembagian Data: Membagi dataset menjadi set data latih (training set) dan set data uji (test set) dengan proporsi 70:30 menggunakan teknik pembagian acak (random split).

A. Algoritma Pembelajaran Mesin

Penelitian ini menggunakan tiga algoritma pembelajaran mesin yang berbeda untuk membangun model prediktif: K-Nearest Neighbors (KNN), Decision Tree, dan Naïve Bayes. Masing-masing algoritma memiliki karakteristik dan pendekatan yang unik dalam memproses data dan membuat prediksi.

1) K-Nearest Neighbors (KNN)
   KNN adalah algoritma non-parametrik yang digunakan untuk klasifikasi dan regresi. Algoritma ini bekerja dengan mencari 'k' titik data terdekat (neighbors) dalam ruang fitur dan menggunakan informasi dari neighbors tersebut untuk membuat prediksi. Implementasi: Dalam penelitian ini, KNN diterapkan dengan menggunakan nilai 'k' yang dipilih berdasarkan validasi silang (cross-validation) untuk memastikan akurasi yang optimal.
2) Decision Tree
   Decision Tree adalah algoritma yang menggunakan struktur pohon untuk membuat keputusan dan prediksi. Pohon keputusan dibangun dengan membagi dataset ke dalam subset berdasarkan fitur yang paling signifikan dalam memisahkan data. Implementasi: Algoritma ini diimplementasikan menggunakan kriteria pemilihan fitur yang memaksimalkan informasi yang diperoleh pada setiap pembagian (split).
3) Naïve Bayes
   Naïve Bayes adalah algoritma berbasis probabilistik yang menggunakan Teorema Bayes dengan asumsi independensi antar fitur. Meskipun sederhana, algoritma ini sering memberikan hasil yang baik terutama pada dataset yang besar. Implementasi: Naïve Bayes diimplementasikan untuk klasifikasi biner dengan mengubah variabel target menjadi biner berdasarkan median dari jumlah kasus TB.

B. Evaluasi Model

Untuk mengevaluasi performa model, digunakan beberapa metrik yang berbeda tergantung pada jenis algoritma yang digunakan:

1) KNN dan Decision Tree (Regresi)
   Mean Absolute Error (MAE): Mengukur rata-rata kesalahan absolut antara nilai yang diprediksi dan nilai sebenarnya. R-squared (R^2): Mengukur proporsi varians dalam variabel dependen yang dapat dijelaskan oleh variabel independen.
2) Naïve Bayes (Klasifikasi)
   Akurasi: Mengukur persentase prediksi yang benar. Confusion Matrix: Matriks yang menunjukkan jumlah true positives, true negatives, false positives, dan false negatives. Classification Report: Menyediakan metrik seperti precision, recall, dan F1-score.

C. Implementasi dan Pengujian

Seluruh proses implementasi dan pengujian dilakukan menggunakan bahasa pemrograman Python dan pustaka pembelajaran mesin seperti scikit-learn. Data diproses dan model dilatih menggunakan fungsi-fungsi yang tersedia dalam pustaka ini. Hasil dari setiap model dievaluasi dan dibandingkan untuk menentukan algoritma yang memberikan performa terbaik dalam prediksi jumlah kasus TB di Indonesia.

III. HASIL DAN PEMBAHASAN

Penelitian ini mengevaluasi performa tiga algoritma pembelajaran mesin yang berbeda, yaitu K-Nearest Neighbors (KNN), Decision Tree, dan Naïve Bayes, dalam memprediksi jumlah kasus Tuberkulosis (TB) di berbagai provinsi di Indonesia. Berikut adalah hasil evaluasi dari masing-masing algoritma.

1) K-Nearest Neighbors (KNN)

```
Mean Absolute Error (MAE) KNN Regressor: 16298.90909090909
R-squared (R²) KNN Regressor: 0.3423926793377091
```

2) Decision Tree

```
Mean Absolute Error (MAE) Decision Tree Regressor: 11948.454545454546
R-squared (R²) Decision Tree Regressor: 0.6246961813794368
```

3) Naïve Bayes

```
Accuracy Naïve Bayes: 1.0
Confusion Matrix Naïve Bayes:
 [[5 0]
 [0 6]]
Classification Report Naïve Bayes:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         5
           1       1.00      1.00      1.00         6

    accuracy                           1.00        11
   macro avg       1.00      1.00      1.00        11
weighted avg       1.00      1.00      1.00        11
```

A. K-Nearest Neighbors (KNN)

Algoritma KNN, yang dikenal karena kesederhanaannya dan kemampuannya dalam menangani data non-linear, memberikan hasil yang menunjukkan kelemahan dalam konteks prediksi jumlah kasus TB. Nilai Mean Absolute Error (MAE) sebesar 16,298.91 menunjukkan bahwa prediksi model ini memiliki deviasi yang cukup besar dari nilai sebenarnya. Hal ini berarti, secara rata-rata, prediksi model ini meleset sekitar 16,298 kasus TB dari nilai aktual.

Selain itu, nilai R-squared (R^2) sebesar 0.3424 menunjukkan bahwa model KNN hanya mampu menjelaskan sekitar 34.24% variansi dalam data. Ini menandakan bahwa model ini tidak mampu menangkap pola dari dataset dengan baik. Salah satu alasan utama dari performa yang kurang memuaskan ini adalah sifat algoritma KNN yang sangat bergantung pada parameter 'k' dan dapat terpengaruh oleh outliers dalam data. Oleh karena itu, meskipun KNN memiliki keunggulan dalam kesederhanaan implementasi, namun dalam kasus ini, model ini tidak cukup kuat untuk memberikan prediksi yang akurat.

B. Decision Tree

Algoritma Decision Tree menunjukkan hasil yang lebih baik dibandingkan dengan KNN. Dengan nilai MAE sebesar 11,948.45, model ini memiliki deviasi yang lebih rendah dari nilai sebenarnya dibandingkan dengan KNN. Hal ini menunjukkan bahwa prediksi dari model Decision Tree lebih dekat dengan nilai aktual, sehingga lebih dapat diandalkan.

Nilai R-squared (R^2) sebesar 0.6247 mengindikasikan bahwa model ini mampu menjelaskan sekitar 62.47% variansi dalam data. Ini menandakan bahwa Decision Tree lebih efektif dalam menangkap pola dalam dataset dibandingkan dengan KNN. Decision Tree memiliki keunggulan dalam kemampuannya untuk menangani data yang memiliki interaksi non-linear antar variabel. Selain itu, struktur pohon keputusan yang dihasilkan memungkinkan interpretasi yang mudah dan visualisasi yang jelas dari proses pengambilan keputusan. Namun, Decision Tree juga memiliki kelemahan seperti overfitting, terutama jika tidak dilakukan pemangkasan (pruning) dengan baik.

C. Naïve Bayes

Model Naïve Bayes menunjukkan performa yang sangat baik dalam klasifikasi dengan akurasi sempurna (1.0). Ini berarti bahwa model ini mampu memprediksi semua instance dalam data uji dengan benar. Confusion Matrix menunjukkan tidak adanya kesalahan prediksi, dengan 5 true positives dan 6 true negatives. Nilai precision, recall, dan f1-score semuanya mencapai 1.0, menunjukkan performa yang sangat baik dalam semua metrik evaluasi.

Keunggulan utama dari Naïve Bayes adalah kesederhanaan dan efisiensinya. Dengan dasar probabilistik, model ini dapat memberikan prediksi yang cepat dan efektif bahkan dengan jumlah data pelatihan yang relatif kecil. Namun, perlu dicatat bahwa hasil akurasi sempurna ini mungkin dipengaruhi oleh ukuran dataset yang relatif kecil. Validitas dari hasil ini perlu diuji lebih lanjut dengan dataset yang lebih besar untuk memastikan bahwa model ini tidak overfitting.

IV. KESIMPULAN

Dari hasil evaluasi, dapat disimpulkan bahwa, model Decision Tree memberikan performa yang lebih baik dalam hal regresi dibandingkan dengan KNN, dengan nilai MAE yang lebih rendah dan R-squared yang lebih tinggi. Decision Tree mampu menangkap pola dalam data dengan lebih baik, menjadikannya pilihan yang lebih cocok untuk prediksi jumlah kasus TB berdasarkan jumlah penduduk. Model Naïve Bayes menunjukkan performa yang sangat baik dalam klasifikasi dengan akurasi

sempurna. Hasil ini menunjukkan potensi besar Naïve Bayes dalam tugas klasifikasi, meskipun validitas hasil perlu diuji lebih lanjut dengan dataset yang lebih besar. Berdasarkan hasil ini, model Decision Tree dapat direkomendasikan untuk prediksi jumlah kasus TB jika pendekatan regresi yang diinginkan. Namun, untuk klasifikasi, Naïve Bayes menunjukkan hasil yang sangat memuaskan dan dapat digunakan untuk tugas klasifikasi dengan akurasi tinggi.

DAFTAR PUSTAKA

[1] Chen, L., et al. (2018). "Machine Learning for Public Health: A Review of Methods and Applications." Annual Review of Public Health, 39: 193-213.
[2] Cover, T., and Hart, P. (1967). "Nearest neighbor pattern classification." IEEE Transactions on Information Theory, 13(1): 21-27.
[3] Perez, L., et al. (2019). "Data Mining in Epidemiology: A Review." Journal of Biomedical Informatics, 94: 103181.
[4] Quinlan, J. R. (1986). "Induction of decision trees." Machine Learning, 1(1): 81-106.
[5] Rish, I. (2001). "An empirical study of the naive Bayes classifier." IJCAI 2001 Workshop on Empirical Methods in AI, 3: 41-46.
[6] World Health Organization. Global Tuberculosis Report 2020. WHO, 2020