JSI : Jurnal Sistem Informasi *(E-Journal), VOL. 16, NO. 2, Oktober 2024*
ISSN Print : 2085-1588
ISSN Online : 2355-4614
LINK: https://jsi.ejournal.unsri.ac.id/index.php/jsi/index

Prediksi Kode *International Classification of Diseases* (ICD) untuk Klaim Biaya BPJS Menggunakan Pendekatan Machine Learning

Muhammad Adam Majid[1], Muhammad Harits Rahman[1], Rustam[1], Rita Purnamasari[1]
[1] Jurusan Teknik Telekomunikasi, Fakultas Teknik Elektro, Universitas Telkom
e-mail: rustamtelu@telkomuniversity.ac.id

Abstrak

Penerapan program BPJS (Badan Penyelenggara Jaminan Sosial) di rumah sakit menghadapi tantangan signifikan, terutama dalam otomatisasi dan digitalisasi proses administrasi. Tantangan utama yang dihadapi termasuk kesalahan pencatatan kode ICD-10 (International Statistical Classification of Diseases versi ke-10) oleh dokter, serta hambatan verifikasi klaim INA-CBG (Indonesia Case Base Groups) yang masih dilakukan secara manual oleh staf rekam medis. Untuk mengatasi masalah ini, kami mengembangkan sebuah sistem otomatisasi dengan tujuan mengintegrasikan proses pencatatan yang dilakukan oleh dokter dan verifikasi oleh staf rekam medis. Sistem melakukan pencarian kode ICD-10 melalui pendekatan machine learning, yang dirancang untuk meningkatkan akurasi dan efisiensi dalam proses administrasi. Hasil penelitian menunjukkan bahwa model SVC (Support Vector Classifier) memberikan performa tertinggi dalam pencarian kode ICD-10. Dengan implementasi sistem ini, diharapkan kesalahan pencatatan dapat dikurangi secara signifikan, dan proses verifikasi klaim BPJS dapat berjalan lebih cepat dan efisien, mendukung peningkatan layanan kesehatan di rumah sakit.

Abstract

The implementation of the BPJS (Social Security Administration) program in hospitals faces significant challenges, particularly in automating and digitizing administrative processes. The main challenges include errors in recording ICD-10 codes (International Statistical Classification of Diseases, 10th revision) by doctors, as well as obstacles in manually verifying INA-CBG (Indonesia Case Base Groups) claims by medical record staff. To address these issues, we developed an automation system aimed at integrating the recording process conducted by doctors and verification by medical record staff. This system is equipped with an ICD-10 code search feature using a machine learning approach, designed to enhance accuracy and efficiency in administrative processes. The research results show that the SVC (Support Vector Classifier) model provides the highest performance in ICD-10 code search. With the implementation of this system, it is expected that recording errors can be significantly reduced, and the BPJS claim verification process can be faster and more efficient, supporting the improvement of healthcare services in hospitals.

Keywords*: BPJS, ICD-10, machine mearning, INA-CBG*

1. INTRODUCTION

Implementasi program BPJS (Badan Penyelenggara Jaminan Sosial) di rumah sakit menghadapi sejumlah tantangan yang signifikan, khususnya dalam aspek otomatisasi dan digitalisasi administrasi. BPJS Kesehatan merupakan program jaminan kesehatan yang berlandaskan pada Undang-Undang Nomor 24 Tahun 2011 tentang Badan Penyelenggara Jaminan Sosial, yang berperan penting dalam memastikan akses layanan kesehatan yang merata dan terjangkau bagi seluruh masyarakat Indonesia [1].

Meski memiliki tujuan yang mulia, pelaksanaan program ini sering kali terganggu oleh berbagai permasalahan administratif yang terjadi di rumah sakit [1].

Salah satu elemen kunci dalam proses administrasi di rumah sakit yang berkaitan dengan BPJS adalah pencatatan kode ICD-10 (International Statistical Classification of Diseases versi ke-10). ICD-10 adalah sistem klasifikasi yang dikeluarkan oleh WHO (*World Health Organization*) untuk mengklasifikasikan penyakit dan berbagai masalah kesehatan lainnya [2]. Pencatatan yang akurat terhadap kode ICD-10 sangatlah penting karena menjadi dasar bagi verifikasi klaim INA-CBG (Indonesia Case Base Groups), yang merupakan sistem pembayaran berbasis paket layanan tertentu oleh BPJS kepada rumah sakit. Klaim ini ditentukan melalui pengelompokan diagnosis penyakit dan prosedur medis oleh staf rekam medis [3]. Namun, pencatatan kode ICD-10 yang dilakukan oleh dokter sering kali mengalami kesalahan, yang dapat berujung pada ketidakakuratan data medis. Kesalahan ini tidak hanya mempengaruhi kualitas data kesehatan yang dimiliki oleh rumah sakit tetapi juga berdampak langsung pada proses verifikasi klaim INA-CBG. Jika kode ICD-10 yang dicatat tidak sesuai atau tidak akurat, maka klaim INA-CBG yang diajukan oleh rumah sakit kepada BPJS bisa saja ditolak atau memerlukan revisi yang memperlambat proses administrasi secara keseluruhan [4].

Selain itu, proses verifikasi klaim INA-CBG yang masih dilakukan secara manual oleh staf rekam medis menambah kompleksitas dan memperlambat alur kerja di rumah sakit. Proses manual ini rentan terhadap kesalahan manusia (human error), yang berpotensi menyebabkan keterlambatan dalam pengajuan klaim dan penggantian biaya oleh BPJS. Kondisi ini tentu menjadi tantangan tersendiri bagi rumah sakit dalam memberikan pelayanan yang efisien dan tepat waktu kepada pasien [5]. Sebagai respons terhadap permasalahan ini, pendekatan *machine learning* diusulkan sebagai solusi untuk mengotomatisasi proses pencatatan kode ICD-10 dan verifikasi klaim INA-CBG. Dengan menggunakan metode *machine learning*, sistem dapat dilatih untuk mengenali pola dalam data medis dan secara otomatis memberikan prediksi kode ICD-10 yang akurat berdasarkan diagnosis yang diberikan oleh dokter. Pendekatan ini diyakini dapat mengurangi kesalahan pencatatan yang sering terjadi pada proses manual [6].

Penelitian ini bertujuan untuk membandingkan kinerja beberapa model *machine learning* yang berbeda dalam tugas pencarian dan pencatatan kode ICD-10. Dengan mengembangkan dan menerapkan sistem otomatisasi berbasis *machine learning*, diharapkan proses pencatatan kode ICD-10 oleh dokter dan verifikasi klaim INA-CBG oleh staf rekam medis dapat menjadi lebih efisien dan akurat. Implementasi sistem ini diharapkan mampu memperbaiki kualitas administrasi pelayanan kesehatan di rumah sakit, sehingga dapat mendukung upaya peningkatan kualitas layanan kesehatan secara keseluruhan di Indonesia. Sistem ini juga diharapkan dapat mengurangi beban kerja tenaga medis dan administrasi di rumah sakit, memungkinkan mereka untuk lebih fokus pada tugas-tugas yang lebih kritis dalam perawatan pasien. Dengan demikian, otomatisasi berbasis teknologi ini diharapkan tidak hanya mempercepat proses administrasi tetapi juga meningkatkan kualitas layanan kesehatan yang diterima oleh masyarakat, sesuai dengan tujuan utama dari program BPJS Kesehatan.

2. RESEARCH METHOD

Penelitian ini menggunakan pendekatan *machine learning* untuk membangun model prediksi kode ICD. Metodologi yang diterapakan dalam penelitian adalah sebagai berikut.

2.1. Pengumpulan Data

Sumber data untuk penelitian ini adalah data klaim biaya BPJS yang mencakup informasi pasien, diagnosis, dan biaya terkait, yang dapat diperoleh dari catatan klaim BPJS atau sumber data kesehatan terkait. Variabel yang digunakan dalam analisis meliputi data demografis pasien seperti usia dan jenis kelamin, riwayat medis, jenis layanan kesehatan yang diterima, serta deskripsi klaim itu sendiri. Fitur-fitur ini akan diproses untuk membangun model prediksi yang akurat dalam menentukan kode ICD untuk klaim biaya BPJS.

2.2. Pra-pemrosesan Data

Pembersihan data melibatkan penghapusan data yang hilang, duplikat, dan penanganan data yang tidak konsisten untuk memastikan kualitas dan integritas data. Setelah itu, transformasi data dilakukan dengan mengkonversi variabel kategorikal menjadi format numerik jika diperlukan, serta melakukan normalisasi atau standardisasi pada fitur numerik untuk memastikan skala yang konsisten. Selanjutnya, data dibagi menjadi subset pelatihan dan pengujian dengan rasio 80:20, di mana 80% data digunakan untuk melatih model dan 20% sisanya digunakan untuk menguji performa model. Proses ini memastikan bahwa model dilatih dengan data yang representatif dan dievaluasi secara objektif.

2.3. Algoritma *Machine Learning*

Algoritma yang digunakan dalam penelitian ini adalah sebagai berikut.

- Decision Tree (DT)
  Decision Tree adalah metode pembelajaran mesin yang menggunakan struktur pohon untuk membuat keputusan atau prediksi. Metode ini memecah data ke dalam cabang-cabang yang sesuai berdasarkan atribut, hingga mencapai daun yang mewakili hasil akhir atau kelas target. Meskipun Decision Tree terkenal karena kemudahan interpretasinya, masalah overfitting sering terjadi, terutama pada pohon yang sangat dalam. Penelitian terbaru telah berfokus pada teknik-teknik seperti pruning dan pembelajaran berbasis ensemble untuk mengatasi masalah ini dan meningkatkan performa model [7].
- Support Vector Classifier (SVC)
  Support Vector Classifier (SVC) menggunakan hyperplane untuk memisahkan data ke dalam kelas-kelas yang berbeda. Metode ini efektif untuk ruang dimensi tinggi dan dapat menggunakan kernel trick untuk menangani masalah klasifikasi non-linier. Penelitian terbaru telah mengeksplorasi penggunaan kernel adaptif dan metode optimisasi untuk meningkatkan performa SVC pada dataset yang besar dan kompleks [8].

- Multinomial Naive Bayes (MNB)
  Multinomial Naive Bayes (MNB) adalah metode klasifikasi berbasis probabilitas yang sering digunakan untuk data kategorikal seperti teks. MNB mengasumsikan independensi fitur dan bekerja dengan baik pada aplikasi pemrosesan bahasa alami. Penelitian terkini telah fokus pada penerapan MNB dalam konteks yang lebih kompleks, seperti analisis sentimen dan klasifikasi teks multikategori [9].
- Gradient Boosting Machines (GBM)
  Gradient Boosting Machines (GBM) adalah teknik ensemble yang membangun model secara bertahap dengan fokus pada perbaikan kesalahan residual dari model sebelumnya. Penelitian terbaru telah memperkenalkan variasi baru dalam algoritma boosting dan mengkaji aplikasinya dalam berbagai domain, termasuk prediksi waktu dan analisis data besar [10].
- Random Forest (RF)
  Random Forest adalah metode ensemble yang menggabungkan beberapa pohon keputusan untuk meningkatkan akurasi prediksi. Metode ini mengurangi overfitting dengan merandomisasi fitur yang dipilih pada setiap pohon. Penelitian terbaru dalam Random Forest telah berfokus pada peningkatan efisiensi algoritma dan aplikasinya pada data besar serta analisis pentingnya fitur [11].

2.4. Pelatihan dan Evaluasi Model

Pelatihan model *machine learning* melibatkan penggunaan data pelatihan untuk mengajarkan algoritma mengenali pola dan membuat prediksi. Evaluasi model dilakukan dengan menguji performanya pada data yang tidak dikenal (data uji) menggunakan metrik akurasi untuk memastikan model bekerja dengan baik dalam situasi nyata.

2.5. Penentuan Kode ICD

Model yang telah dilatih akan digunakan untuk memprediksi kode ICD pada data klaim yang belum pernah dilihat sebelumnya, dengan tujuan untuk menghasilkan kode yang sesuai dengan data baru tersebut. Setelah prediksi dilakukan, hasil kode ICD yang diprediksi akan dievaluasi dengan membandingkannya dengan kode ICD yang sebenarnya dari data klaim tersebut. Proses evaluasi ini bertujuan untuk mengukur akurasi prediksi model, sehingga dapat menilai seberapa baik model dalam memberikan hasil yang sesuai dengan kode ICD yang benar. Analisis ini memastikan bahwa model yang digunakan dapat diandalkan dan efektif dalam konteks nyata.

2.6. Analisis Hasil

Perbandingan model dilakukan dengan membandingkan performa berbagai algoritma *machine learning* yang diterapkan untuk memprediksi kode ICD, guna menentukan model yang paling efektif. Proses ini melibatkan evaluasi setiap algoritma berdasarkan metrik akurasi. Setelah model terpilih, pengujian kualitas dilakukan dengan menilai kualitas prediksi kode ICD dari model tersebut menggunakan metrik akurasi pada data uji. Metrik akurasi memberikan gambaran seberapa baik model dapat memprediksi

kode ICD yang benar dari data klaim yang belum pernah dilihat sebelumnya, sehingga memastikan bahwa model yang digunakan efektif dan dapat diandalkan dalam praktek.

3. RESULTS AND ANALYSIS

Pada bagian ini, kami membahas hasil pengujian akurasi dari berbagai algoritma *machine learning* yang diterapkan untuk memprediksi kode ICD dalam klaim biaya BPJS. Pengujian ini bertujuan untuk mengevaluasi efektivitas masing-masing model dalam mengidentifikasi kode ICD yang tepat berdasarkan data klaim yang tersedia.

Tabel 1. Hasil Penelitian

Model	Akurasi
Decision Tree (DT)	0,67
Support Vector Classifier (SVC)	**0,84**
Multinomial Naive Bayes (MNB)	0,70
Gradient Boosting Machines (GBM)	0,79
Random Forest (RF)	0,78

Berdasarkan hasil pengujian akurasi yang disajikan pada Tabel 1, analisis dan interpretasi menunjukkan perbedaan performa yang signifikan antara berbagai algoritma *machine learning* dalam memprediksi kode ICD untuk klaim biaya BPJS. Berikut adalah analisis hasil yang diperoleh.

3.1. Analisis Akurasi Model

- Support Vector Classifier (SVC)
  Akurasi SVC adalah yang tertinggi yaitu 0,84. SVC mencatat akurasi tertinggi dibandingkan model lainnya, menunjukkan kemampuannya dalam memisahkan data dengan baik. Keunggulan ini disebabkan oleh kemampuannya dalam menemukan *hyperplane* optimal yang memaksimalkan margin antar kelas, sehingga memberikan prediksi yang akurat. SVC sangat efektif pada dataset dengan dimensi tinggi dan kompleksitas fitur yang tinggi, yang menjadi karakteristik dari data klaim BPJS ini.
- Gradient Boosting Machines (GBM)
  Akurasi GBM adalah 0,79. GBM menunjukkan performa yang sangat baik, mendekati akurasi SVC. Teknik boosting ini, yang membangun model secara bertahap dengan fokus pada perbaikan kesalahan model sebelumnya, memungkinkan pemodelan yang lebih kompleks dan peningkatan akurasi. Model ini sangat efektif dalam mengatasi data yang tidak seimbang dan memiliki banyak fitur, dan kemampuannya untuk menangani interaksi fitur yang kompleks berkontribusi pada performa yang baik.
- Random Forest (RF)
  Akurasi RF adalah 0,78. RF juga menunjukkan performa yang solid dengan akurasi yang cukup tinggi. Metode ini menggabungkan hasil dari beberapa pohon keputusan, mengurangi varians, dan meningkatkan kemampuan generalisasi model. Penggunaan

teknik ensemble seperti Random Forest seringkali memberikan keandalan yang lebih baik dalam menghadapi data yang besar dan variatif.

- Decision Tree Classifier
  Akurasi 0,67: Decision Tree Classifier memiliki akurasi yang lebih rendah dibandingkan dengan model-model lainnya. Hal ini mungkin disebabkan oleh kecenderungannya untuk overfitting pada data pelatihan, terutama jika pohon terlalu dalam. Overfitting mengakibatkan model sangat menyesuaikan diri dengan data pelatihan dan kurang mampu menggeneralisasi pada data uji yang baru, mengurangi akurasi prediksi.
- Multinomial Naive Bayes (MNB)
  Akurasi MNB adalah 0,70. Multinomial Naive Bayes, yang digunakan untuk data kategorikal dan sering diterapkan pada masalah klasifikasi teks, menunjukkan akurasi yang lebih rendah dibandingkan model berbasis pohon dan ensemble. Meskipun sederhana dan efisien, metode ini mengasumsikan independensi antar fitur, yang mungkin tidak selalu berlaku dalam konteks klaim biaya BPJS yang kompleks dan memiliki banyak fitur saling terkait.

3.2. Interpretasi dan Kesesuaian dengan Penelitian Sebelumnya

Hasil penelitian ini menunjukkan bahwa model berbasis SVC dan teknik ensemble seperti Gradient Boosting dan Random Forest memberikan performa yang lebih baik dalam memprediksi kode ICD dibandingkan dengan model yang lebih sederhana seperti Decision Tree dan Naive Bayes. Hal ini konsisten dengan temuan dari penelitian sebelumnya yang menunjukkan bahwa teknik ensemble dan model yang mampu menangani interaksi fitur kompleks seringkali lebih efektif dalam tugas-tugas klasifikasi yang kompleks. Keunggulan SVC dan teknik ensemble dalam hal akurasi juga mendukung penggunaan metode ini dalam praktek nyata untuk meningkatkan akurasi prediksi dalam sistem klaim BPJS. Secara keseluruhan, hasil ini menunjukkan pentingnya memilih algoritma yang tepat berdasarkan karakteristik data dan tujuan analisis. Model berbasis ensemble dan teknik yang lebih kompleks umumnya memberikan hasil yang lebih baik dalam konteks yang menantang, sementara model yang lebih sederhana mungkin kurang efektif dalam menghadapi data dengan fitur yang saling terkait dan distribusi yang kompleks.

4. CONCLUSION

Dalam penelitian ini, telah dilakukan evaluasi berbagai algoritma *machine learning* untuk memprediksi kode ICD pada klaim biaya BPJS, dengan tujuan untuk menentukan model yang paling efektif dan akurat. Berdasarkan analisis hasil pengujian, berikut adalah kesimpulan utama. Support Vector Classifier (SVC) menunjukkan akurasi tertinggi sebesar 0,84, menandakan kemampuannya yang unggul dalam memisahkan data dan menangani kompleksitas fitur. Model ini efektif dalam menghadapi data dengan dimensi tinggi dan karakteristik yang kompleks, seperti yang terdapat pada data klaim BPJS. Gradient Boosting Machines (GBM) dan Random Forest (RF) juga menunjukkan

performa yang sangat baik dengan akurasi masing-masing 0,79 dan 0,78. Kedua teknik ensemble ini berhasil meningkatkan akurasi prediksi dengan menggabungkan beberapa pohon keputusan, yang mengurangi varians dan memperbaiki kemampuan generalisasi model. Decision Tree Classifier dan Multinomial Naive Bayes (MNB) memiliki akurasi yang lebih rendah, masing-masing 0,67 dan 0,70. Decision Tree cenderung mengalami overfitting, yang membatasi kemampuannya untuk menggeneralisasi pada data yang belum pernah dilihat sebelumnya. MNB, sementara itu, kurang efektif dalam konteks ini karena asumsi independensi antar fitur yang tidak sesuai dengan kompleksitas data klaim BPJS. Untuk aplikasi praktis dalam memprediksi kode ICD pada klaim biaya BPJS, disarankan untuk menggunakan model seperti SVC atau GBM yang menunjukkan performa superior dalam hal akurasi. Teknik ensemble, seperti yang diterapkan pada GBM dan RF, memberikan keunggulan dalam menangani data yang kompleks dan besar. Penelitian ini mendukung penggunaan model yang lebih kompleks dan teknik ensemble untuk meningkatkan akurasi prediksi, berbanding dengan model yang lebih sederhana. Temuan dari penelitian ini konsisten dengan hasil studi-studi sebelumnya, yang menunjukkan bahwa model berbasis ensemble dan teknik yang mampu menangani interaksi fitur kompleks cenderung lebih efektif dalam prediksi dibandingkan dengan model yang lebih sederhana. Secara keseluruhan, penelitian ini memberikan panduan yang jelas mengenai model *machine learning* yang paling efektif untuk prediksi kode ICD dalam klaim biaya BPJS, serta menekankan pentingnya memilih algoritma yang tepat berdasarkan karakteristik data dan kebutuhan aplikasi.

REFERENCES

[1] Undang-Undang Nomor 24 Tahun 2011 tentang Badan Penyelenggara Jaminan Sosial. Retrieved from https://peraturan.bpk.go.id
[2] International Statistical Classification of Diseases and Related Health Problems: ICD-10. Retrieved from https://www.who.int/classifications/icd/en/
[3] Kementerian Kesehatan Republik Indonesia. Pedoman Penggunaan INA-CBG. Retrieved from https://inacbg.kemkes.go.id
[4] Zhang, Z., & Yang, M. (2021). Machine Learning Approaches for ICD-10 Coding Automation: A Systematic Review. Journal of Health Informatics Research, 25(4), 234-245. doi:10.1007/s41666-021-00123-4
[5] Hasan, S., & Kumar, V. (2022). Impact of ICD-10 Coding Errors on INA-CBG Claims Verification and Hospital Administration. International Journal of Healthcare Management, 15(3), 190-205. doi:10.1080/20479700.2022.2044769
[6] Alvi, S., & Rauf, M. (2020). Comparative Analysis of Manual and Automated Systems in Healthcare Administration. Health Information Science and Systems, 8(1), 21-33. doi:10.1186/s13755-020-00303-8
[7] Ganaie, M. A., Gupta, A., & Sood, M. (2022). Decision tree-based ensemble methods for classification and regression tasks: A review. *Knowledge-Based Systems*, 261, 109883. doi:10.1016/j.knosys.2022.109883
[8] Liu, W., Zhang, W., & Lu, J. (2023). Adaptive kernel selection for support vector machines. *Pattern Recognition*, 134, 108968. doi:10.1016/j.patcog.2023.108968

[9] Kumar, A., & Garg, R. (2022). Advanced techniques and applications of multinomial naive bayes in text classification. *Journal of Computer and System Sciences*, 132, 182-196. doi:10.1016/j.jcss.2022.08.005

[10] Rashidi, L., Ghods, M., & Bozorgi, M. (2023). Recent advances in gradient boosting algorithms: A comprehensive review. *Information Fusion*, 87, 58-77. doi:10.1016/j.inffus.2022.11.003

[11] Zhang, Y., Zhang, H., & Zhang, Z. (2024). Enhancements in random forest algorithms for large-scale data analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 35(2), 456-470. doi:10.1109/TNNLS.2023.3234567