Stroke Prediction using Machine Learning

Yusra Erlangga Putra[1], Sheryl Anastasya[2], Resky Auliyah Kartini Askin[3], Ivan Betrandi[4], Amaliah Diah[5]
Departemen Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Hasanuddin, Makassar, Indonesia
[1]yusraerlangg@gmail.com, [2]jane.doe@email.com

Abstrak

Stroke merupakan penyebab utama kecacatan jangka panjang dan kematian di seluruh dunia, dengan risiko yang meningkat seiring dengan bertambahnya usia serta adanya faktor risiko seperti hipertensi dan diabetes. Tujuan penelitian ini adalah mengembangkan model prediksi dini yang akurat untuk stroke sehingga memungkinkan intervensi perawatan kesehatan preventif yang efektif. Penelitian ini menggunakan algoritma Random Forest dan Support Vector Machine (SVM). Karena ketidakseimbangan pada himpunan data, teknik Synthetic Minority Oversampling Technique (SMOTE) diterapkan untuk meningkatkan representasi data minoritas. Model dioptimalkan melalui penyesuaian hiperparameter menggunakan Optimasi Bayesian. Evaluasi model dilakukan dengan metrik akurasi, presisi, sensitivitas, dan skor-F1, dengan validasi silang untuk memastikan keandalan pada data yang belum terlihat. Hasil eksperimen menunjukkan bahwa algoritma SVM dengan SMOTE dan Optimasi Bayesian mencapai akurasi tertinggi. Temuan ini menunjukkan bahwa model pembelajaran mesin yang dioptimalkan dapat memberikan kontribusi signifikan dalam prediksi dini stroke dan mendukung pengambilan keputusan dalam sistem perawatan kesehatan preventif.

Kata kunci: *Prediksi Stroke, pembelajaran mesin, Random Forest, Support Vector Machine, SMOTE, Bayesian Optimization*

1. Introduction

1.1 Latar Belakang

Stroke merupakan masalah kesehatan global yang serius, menempati posisi kedua sebagai penyebab kematian tertinggi dan posisi ketiga sebagai penyebab kecacatan di dunia. Menurut data WHO, satu dari empat orang berisiko mengalami stroke dalam masa hidupnya, dengan 70% kasus terjadi di negara-negara berpenghasilan rendah dan menengah[1]. Kompleksitas faktor risiko stroke yang meliputi tekanan darah tinggi, kolesterol, diabetes, obesitas, dan gaya hidup, menjadikan prediksi stroke sebagai tantangan yang signifikan dalam dunia kesehatan.[2]

Prediksi dini stroke menggunakan metode konvensional seringkali terbatas dalam kemampuannya mengintegrasikan berbagai faktor risiko secara simultan. Perkembangan pembelajaran mesin membuka peluang baru dalam meningkatkan akurasi prediksi stroke. Dengan kemampuannya menganalisis data kompleks dalam jumlah besar, teknik pembelajaran mesin dapat mengidentifikasi pola-pola tersembunyi dan korelasi antar berbagai faktor risiko yang sulit dideteksi melalui metode konvensional.

Penelitian ini mengusulkan pendekatan inovatif dengan mengombinasikan algoritma Random Forest dan Support Vector Machine (SVM) untuk prediksi stroke. Kedua algoritma ini dipilih karena kemampuannya dalam menangani data kesehatan yang kompleks dan menghasilkan model prediksi yang akurat. Untuk mengatasi masalah ketidakseimbangan data yang umum dalam kasus medis, diterapkan teknik SMOTE (Synthetic Minority Oversampling Technique). Selanjutnya, optimasi Bayesian digunakan untuk meningkatkan performa model melalui penyetelan hiperparameter yang optimal. Pendekatan komprehensif ini diharapkan dapat menghasilkan sistem prediksi stroke yang lebih andal dan aplikatif dalam praktik klinis.

1.2 Literature Review

1.3 Research Rationale

1.4 Research Questions and Objectives

Penelitian ini didasari oleh beberapa pertanyaan penelitian yang berkaitan dengan keefektifan algoritma Random Forest dan Support Vector Machine dalam memprediksi risiko stroke berdasarkan berbagai faktor kesehatan. Selain itu, penelitian ini juga mengkaji sejauh mana penerapan teknik SMOTE dapat meningkatkan kinerja model dalam menangani ketidakseimbangan data pada kasus prediksi stroke, serta bagaimana Optimasi Bayesian dapat memengaruhi kinerja model dalam hal akurasi dan presisi prediksi stroke.

Berdasarkan pertanyaan penelitian tersebut, penelitian ini bertujuan untuk mengembangkan dan membandingkan model prediksi stroke menggunakan algoritma Random Forest dan Support Vector Machine untuk mengidentifikasi pendekatan yang paling efektif. Penelitian ini juga akan mengevaluasi dampak penerapan SMOTE dalam meningkatkan kualitas prediksi pada kasus dengan sebaran data yang tidak seimbang, serta mengoptimalkan parameter model menggunakan Optimasi Bayesian untuk mencapai kinerja prediksi yang optimal. Tujuan akhir penelitian ini adalah menghasilkan model prediksi stroke yang dapat diterapkan dalam sistem pendukung keputusan klinis untuk deteksi dini risiko stroke.

2. Research Methods

2.1 Dataset Description

Dataset yang digunakan dalam penelitian ini berasal dari *Kaggle Stroke Prediction Dataset*[3], yang berisi

informasi kesehatan dari 5110 pasien. Dataset ini mencakup berbagai parameter demografis dan faktor risiko kesehatan yang berpotensi terkait dengan kejadian stroke. Data dikumpulkan dari berbagai fasilitas kesehatan dan mencakup informasi seperti usia, jenis kelamin, berbagai penyakit, dan gaya hidup pasien.

Table 1: Deskripsi Fitur Dataset Stroke Prediction

Fitur	Deskripsi	Tipe Data
id	Identifier unik setiap pasien	Numerik
gender	Jenis kelamin pasien (*Male*, *Female*, *Other*)	Kategorikal
age	Usia pasien	Numerik
hypertension	0 jika pasien tidak memiliki hipertensi, 1 jika memiliki hipertensi	Biner
heart_disease	0 jika pasien tidak memiliki penyakit jantung, 1 jika memiliki penyakit jantung	Biner
ever_married	Status pernikahan (*Yes*, *No*)	Kategorikal
work_type	Tipe pekerjaan (*children*, *Govt_job*, *Never_worked*, *Private*, *Self-employed*)	Kategorikal
Residence_type	Tipe tempat tinggal (*Rural*, *Urban*)	Kategorikal
avg_glucose_level	Level glukosa rata-rata dalam darah (*Average Glucose Level*)	Numerik
bmi	Body Mass Index	Numerik
smoking_status	Status merokok (*formerly smoked*, *never smoked*, *smokes*, *Unknown*)	Kategorikal
stroke	1 jika pasien pernah stroke, 0 jika tidak	Biner

2.2 Performance Metrics

Dalam evaluasi model klasifikasi stroke, kami menggunakan beberapa metrik performa standar yang diperoleh dari *confusion matrix*.

Accuracy mengukur proporsi total prediksi yang benar dibandingkan dengan semua kasus.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision mengukur proporsi prediksi positif yang benar-benar positif.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall (Sensitivity) mengukur proporsi kasus positif yang berhasil diidentifikasi.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F1-Score merupakan rata-rata harmonik dari *precision* dan *recall*.

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Dalam konteks ini, TP (*True Positive*) adalah jumlah kasus stroke yang diprediksi benar sebagai stroke. TN (*True Negative*) merupakan jumlah kasus non-stroke yang diprediksi benar sebagai non-stroke. FP (*False Positive*) menunjukkan jumlah kasus non-stroke yang salah diprediksi sebagai stroke. FN (*False Negative*) adalah jumlah kasus stroke yang salah diprediksi sebagai non-stroke.

Pemilihan metrik-metrik ini didasarkan pada karakteristik dataset yang tidak seimbang, sehingga F1-score menjadi sangat penting karena memberikan gambaran yang lebih baik tentang performa model pada kasus yang tidak seimbang dibandingkan dengan *accuracy* saja.

2.3 Formulas

$$E = mc^2 \tag{5}$$

3. Results and Discussion

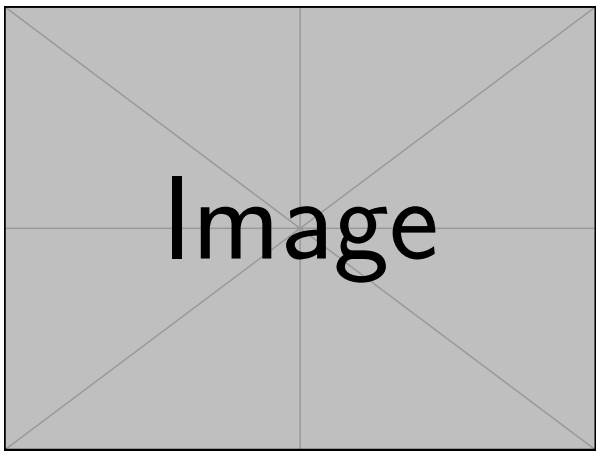

Figure 1: A single-column figure example.

Table 2: Dataset characteristics for stroke prediction

Characteristic	Value	Percentage
Total Samples	5110	100%
Stroke Cases	249	4.87%
Non-stroke Cases	4861	95.13%

4. Conclusion

Acknowledgements

References

[1] World Health Organization and Dr. Poonam Khetrapal Singh, WHO Regional Director for South-East

Asia, “World stroke day,” 2021. Accessed: 27 November 2024.

[2] M. U. Emon, M. S. Keya, T. I. Meghla, M. M. Rahman, M. S. Al Mamun, and M. S. Kaiser, “Performance analysis of machine learning approaches in stroke prediction,” in *2020 4th international conference on electronics, communication and aerospace technology (ICECA)*, pp. 1464–1469, IEEE, 2020.

[3] F. Soriano, “Stroke prediction dataset,” 2024. Accessed: 15 Oktober 2024.