

# *Analisis Prediksi Risiko Stroke Menggunakan Algoritma Random Forest Classifier (RFC) Dan Support Vector Machine (SVM)*

Rosmeilina Anggreini

Program Studi : Sistem Informasi, Universitas Darwan Ali

Email : [rosmeilinaanggreini@gmail.com](mailto:rosmeilinaanggreini@gmail.com)

**ABSTRACT**— Stroke is one of the leading causes of death and disability worldwide, making early detection of this disease crucial in efforts to prevent it. This study aims to predict the risk of stroke using two classification algorithms, namely Random Forest Classifier (RFC) and Support Vector Machine (SVM), and to compare their performance. The data processing stages include handling missing values, feature transformation, normalization, and data balancing using SMOTE. The model was trained and tested, then evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics. The results show that RFC has a higher accuracy of 88.45%, but the recall obtained is relatively low at 20%, while SVM has a better recall of 70% with an accuracy of 75.44%. Although the AUC values of both models are not too different, around 0.78 to 0.79, SVM is more suitable for implementation in early stroke detection systems because it has better sensitivity to positive cases.

**Keywords**— *Stroke, Classification, Random Forest Classifier, Support Vector Machine.*

**ABSTRAK**— Stroke adalah salah satu penyebab utama kematian dan kecacatan di dunia, sehingga mendeteksi penyakit ini sejak dini sangat penting dalam upaya mencegahnya. Penelitian ini bertujuan untuk memprediksi risiko stroke menggunakan dua algoritma klasifikasi, yaitu Random Forest Classifier (RFC) dan Support Vector Machine (SVM), serta membandingkan performanya. Tahapan pengolahan data meliputi penanganan nilai kosong, transformasi fitur, normalisasi, serta penyeimbangan data menggunakan SMOTE. Model dilatih dan diuji, kemudian dievaluasi menggunakan metrik akurasi, precision, recall, F1-score, dan ROC-AUC. Hasil menunjukkan bahwa RFC memiliki akurasi yang lebih tinggi yaitu 88,45%, namun recall yang diperoleh relatif rendah yaitu 20%, sedangkan SVM memiliki recall yang lebih baik yaitu 70% dengan akurasi sebesar 75,44%. Meskipun nilai AUC dari kedua model tidak terlalu jauh berbeda, sekitar 0,78 hingga 0,79, SVM lebih sesuai diterapkan dalam sistem deteksi dini stroke karena memiliki sensitivitas yang lebih baik terhadap kasus positif.

**Kata kunci**— *Stroke, Klasifikasi, Random Forest Classifier, Support Vector Machine.*

## I. PENDAHULUAN

Stroke adalah salah satu penyakit yang menjadi penyebab utama kematian dan kecacatan di seluruh dunia. Penyakit ini terjadi karena terganggunya aliran darah ke otak, yang disebabkan karena adanya pembekuan darah yang menghalangi aliran darah (jenis iskemik) atau karena pembuluh darah di otak pecah (jenis hemoragik). Saat serangan terjadi, sel-sel otak mulai rusak dan mati hanya dalam beberapa menit. Kondisi ini seringkali menyebabkan dampak jangka panjang berupa kecacatan fisik maupun kemampuan berpikir. Oleh karena itu, upaya untuk mengidentifikasi faktor-faktor risiko yang berkontribusi terhadap stroke sangat penting untuk pencegahan dini dan pengelolaan kesehatan masyarakat.

Dengan berkembangnya teknologi di bidang kecerdasan buatan, pemanfaatan metode prediksi berbasis machine learning dalam dunia medis semakin meluas, terutama dalam upaya deteksi dini penyakit seperti stroke. Penelitian ini bertujuan untuk merancang model prediksi risiko stroke menggunakan dua algoritma klasifikasi, yaitu Random Forest Classifier (RFC) dan Support Vector Machine (SVM). Kedua algoritma tersebut dipilih karena dinilai efektif dalam mengatasi kompleksitas data medis dan mampu memberikan hasil klasifikasi yang akurat. Penelitian sebelumnya, seperti

[1] dan [2], telah menunjukkan bahwa kedua algoritma ini memiliki performa yang baik dalam mengklasifikasikan data kesehatan. Dengan demikian, penelitian ini bertujuan untuk membandingkan performa RFC dan SVM agar dapat menentukan model mana yang paling optimal dalam mendukung pengambilan keputusan medis.

Berdasarkan laporan dari World Stroke Organization (WSO), stroke menempati posisi kedua sebagai penyebab kematian terbanyak di dunia, serta menjadi penyebab kecacatan ketiga, dengan tren peningkatan beban penyakit yang konsisten sejak tahun 1990 hingga saat ini. Di Indonesia, data dari Survei Kesehatan Indonesia (SKI) 2023 menunjukkan bahwa prevalensi stroke mencapai 8,3 per 1.000 penduduk, dengan nilai tertinggi terjadi pada kelompok usia 75 tahun ke atas, yaitu sebesar 41,3 per mil. Faktor risiko utama yang berkontribusi terhadap kejadian stroke antara lain adalah hipertensi, diabetes, obesitas, gangguan jantung, kebiasaan merokok, serta gaya hidup yang kurang sehat. Tingginya angka kejadian stroke dan dampak jangka panjang yang ditimbulkan menekankan perlunya pengembangan sistem deteksi dini berbasis data sebagai bagian dari strategi pencegahan yang lebih efektif dan untuk mendukung pengambilan keputusan medis secara cepat dan tepat.

Walaupun stroke termasuk penyakit yang dapat dicegah, upaya deteksi dini di Indonesia masih menghadapi berbagai tantangan, terutama akibat keterbatasan infrastruktur dan jumlah tenaga kesehatan, terutama di wilayah dengan akses layanan medis yang terbatas. Sebagian besar fasilitas pelayanan kesehatan tingkat pertama belum memiliki sistem digital yang mampu mendukung proses skrining secara otomatis. Sementara itu, metode deteksi yang masih bersifat manual umumnya bergantung pada pemeriksaan klinis dan laboratorium, yang tidak hanya memerlukan biaya besar dan waktu yang lama, tetapi juga keterlibatan langsung tenaga medis. Belum tersedianya sistem prediktif berbasis data yang memadai menjadi hambatan tersendiri. Oleh karena itu, pemanfaatan teknologi menjadi alternatif yang perlu dipertimbangkan untuk meningkatkan efektivitas deteksi dini stroke [3].

Beberapa penelitian telah membuktikan efektivitas algoritma machine learning dalam klasifikasi data medis, terutama dalam prediksi penyakit stroke. [1] menyatakan bahwa algoritma Random Forest Classifier (RFC) mampu mencapai akurasi sebesar 95% dengan nilai AUC 0,80 dalam melakukan prediksi risiko stroke. Di sisi lain, penelitian oleh [2] menunjukkan bahwa algoritma Support Vector Machine (SVM) menggunakan kernel RBF dan metode cross-validation mampu mencapai akurasi sebesar 94,8% serta F1-score sebesar 92,3%. [4] dan [5] juga menunjukkan bahwa SVM memiliki performa yang lebih baik dibandingkan algoritma lain seperti K-Nearest Neighbor (K-NN) dan Decision Tree, meskipun hasil akhir tetap tergantung pada karakteristik dan distribusi data yang digunakan.

Untuk mengatasi permasalahan ketidakseimbangan data, [6] dan [7] menggunakan teknik Synthetic Minority Over-sampling Technique (SMOTE) dan berhasil meningkatkan akurasi model Random Forest hingga 95,5%. Sementara itu, [8] menerapkan metode ADASYN sebelum membangun model SVM dan mencatat peningkatan akurasi yang signifikan. Dari aspek optimasi, [9] menekankan pentingnya penyesuaian hyperparameter, seperti nilai  $n\_estimators$  dan  $max\_depth$ , berguna untuk meningkatkan kinerja algoritma RFC dan Decision Tree. Selain itu [10] menunjukkan bahwa optimasi pada tahap pra-pemrosesan data, termasuk normalisasi dan pengkodean fitur, berpengaruh besar terhadap akurasi model RFC dalam klasifikasi stroke. Adapun efektivitas algoritma RFC dalam klasifikasi data medis juga diperlihatkan oleh [11] melalui penerapan RFC pada klasifikasi penyakit jantung, dengan capaian F1-score sebesar 0,868. Berdasarkan temuan-temuan tersebut, baik RFC maupun SVM memiliki potensi yang kuat untuk diterapkan sebagai model prediksi risiko penyakit kronis berbasis data.

Meskipun penelitian terdahulu menunjukkan performa tinggi dari algoritma Random Forest Classifier (RFC) dan Support Vector Machine (SVM), sebagian besar penelitian tersebut hanya berfokus pada penerapan satu algoritma secara terpisah dan belum melakukan perbandingan langsung. Selain itu, permasalahan seperti

ketidakseimbangan data dan keterbatasan dalam evaluasi menyeluruh masih menjadi tantangan yang belum sepenuhnya diatasi. Penelitian ini bertujuan mengatasi kekurangan tersebut dengan langsung membandingkan hasil RFC dan SVM menggunakan dataset yang sama. Hasil penelitian juga dinilai berdasarkan beberapa metrik seperti akurasi, presisi, recall, dan F1-score. Dengan pendekatan ini, diharapkan bisa dibuat model prediksi risiko stroke yang lebih akurat dan bisa digunakan secara efektif dalam penerapan medis.

Pemilihan tema ini dilatarbelakangi oleh tingginya angka kasus stroke di Indonesia serta belum maksimalnya pemanfaatan teknologi prediktif dalam sistem layanan kesehatan. Dengan menerapkan algoritma Random Forest Classifier (RFC) dan Support Vector Machine (SVM), penelitian ini diharapkan dapat memberikan solusi inovatif dalam mendeteksi stroke secara dini. Selain itu, hasil penelitian ini diharapkan juga dapat membantu proses pengambilan keputusan medis menjadi lebih cepat, efisien, dan didasarkan pada data, sehingga mampu meningkatkan kualitas layanan kesehatan secara keseluruhan.

## II. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan menggunakan data sekunder dalam bentuk eksperimen. Tujuan utama dari penelitian ini adalah membandingkan performa algoritma Random Forest Classifier (RFC) dan Support Vector Machine (SVM) dalam memprediksi kemungkinan seseorang mengalami stroke berdasarkan data yang diperoleh dari pasien. Metodologi penelitian ini terdiri dari beberapa tahapan sebagai berikut:

### A. Pengumpulan Data

Pengumpulan data adalah tahap awal dalam proses penelitian yang bertujuan mengumpulkan informasi yang digunakan untuk analisis dan pembuatan model. Data yang dikumpulkan harus sesuai dengan tujuan penelitian, relevan terhadap isu yang dibahas, serta bisa diproses secara terstruktur menggunakan metode yang sudah ditentukan. Dalam penelitian ini, data diperoleh dari dataset publik yang tersedia di platform Kaggle dengan judul "Stroke Prediction Dataset".

Dataset ini berisi 5.110 data pasien yang berisi informasi tentang usia, jenis kelamin, riwayat hipertensi, penyakit jantung, status pernikahan, jenis pekerjaan, tipe tempat tinggal, kadar glukosa rata-rata, indeks massa tubuh (BMI), dan status merokok. Informasi tersebut digunakan untuk melihat risiko seseorang mengalami stroke. Atribut stroke digunakan sebagai label klasifikasi, di mana nilai 1 menunjukkan pasien pernah mengalami stroke, sedangkan nilai 0 berarti pasien tidak stroke.

Dataset ini dipilih karena tersedia secara gratis, memuat informasi yang lengkap, dan sudah digunakan dalam penelitian sebelumnya. Selain itu, data ini memiliki fitur-fitur yang relevan dalam menentukan risiko terkena stroke, sehingga sangat sesuai digunakan untuk menguji algoritma berbasis machine learning.

## B. Pra-pemrosesan Data

Sebelum data digunakan untuk melatih model machine learning, dilakukan proses pra-pemrosesan data agar dataset menjadi siap dan tidak menghasilkan kesalahan saat pemodelan. Kolom yang memiliki nilai kosong diisi dengan rata-rata dari kolom itu sendiri agar data lebih lengkap. Fitur kategorikal seperti gender, ever\_married, work\_type, residence\_type, dan smoking\_status diubah ke bentuk numerik menggunakan metode label encoding atau one-hot encoding, sehingga dapat dikenali oleh algoritma klasifikasi.

Selanjutnya, fitur numerik seperti age, avg\_glucose\_level, dan bmi dinormalisasi menggunakan StandardScaler, sehingga setiap variabel memiliki nilai rata-rata 0 dan deviasi standar 1. Normalisasi ini penting untuk menjaga kesetaraan skala antar fitur, terutama pada algoritma seperti SVM yang sensitif terhadap perbedaan skala. Setelah itu, data dibagi menjadi data latih sebesar 80% dan data uji sebesar 20% dengan menggunakan metode stratified split untuk memastikan proporsi kelas target tetap seimbang.

Untuk mengatasi ketidakseimbangan kelas, diterapkan metode SMOTE (Synthetic Minority Over-sampling Technique) pada data latih, yang menghasilkan sampel sintetis pada kelas minoritas (pasien stroke). Proses ini membantu menciptakan distribusi kelas yang lebih seimbang, sehingga mengurangi bias terhadap kelas mayoritas dan meningkatkan akurasi serta keadilan model.

## C. Pemodelan Algoritma

Pada tahap ini, penelitian menggunakan dua algoritma klasifikasi untuk membangun model prediksi risiko stroke, yaitu Random Forest Classifier (RFC) dan Support Vector Machine (SVM). Kedua algoritma tersebut dipilih karena telah banyak digunakan dalam penelitian sebelumnya dan terbukti efektif dalam menyelesaikan berbagai permasalahan klasifikasi pada data medis. Penerapan model dilakukan pada data yang telah melewati proses pra-pemrosesan serta penyeimbangan kelas menggunakan metode SMOTE. Tujuan dari proses ini adalah untuk membandingkan performa antara RFC dan SVM dalam memprediksi kemungkinan terjadinya stroke pada individu, berdasarkan atribut-atribut demografis dan klinis yang tersedia dalam dataset.

### 1. Random Forest Classifier (RFC)

Random Forest adalah algoritma machine learning yang menggabungkan beberapa pohon keputusan (decision trees) untuk meningkatkan akurasi dan mengurangi overfitting. Model ini dilatih menggunakan data pelatihan yang telah diseimbangkan dengan metode SMOTE. Proses pelatihan dilakukan menggunakan metode pelatihan (fit) dan prediksi (predict) yang disediakan oleh library pembelajaran mesin.

### 2. Support Vector Machine (SVM)

SVM adalah algoritma untuk mengklasifikasikan data dengan cara mencari garis pemisah (hyperlane) terbaik antar kelas. Dalam penelitian ini, algoritma SVM menggunakan kernel Radial Basis Function (RBF) untuk menangani data yang memiliki pola tidak lurus, dan diatur

dengan opsi probability=True agar memungkinkan perhitungan nilai ROC-AUC. Seperti halnya dengan RFC, model SVM dilatih menggunakan data hasil oversampling SMOTE, dan hasil prediksinya digunakan dalam proses evaluasi performa model.

## D. Evaluasi Model

Bagian ini menjelaskan tahapan evaluasi dalam proses penelitian secara sistematis. Setelah melalui proses pengumpulan dan pra-pemrosesan data, model dibangun dan dilatih menggunakan data latih yang telah diseimbangkan dengan metode SMOTE. Tahapan pelatihan dilakukan secara terpisah untuk dua algoritma yang digunakan, yaitu Random Forest Classifier (RFC) dan Support Vector Machine (SVM). Setelah model selesai dilatih, masing-masing diuji menggunakan data uji yang telah dipisahkan sebelumnya.

Tahapan selanjutnya adalah mengevaluasi performa model berdasarkan hasil prediksi terhadap data uji. Evaluasi dilakukan dengan menghitung beberapa metrik umum, seperti akurasi, presisi, recall, F1-score, dan ROC-AUC. Tujuannya adalah untuk menilai seberapa efektif model dalam mengklasifikasikan data secara akurat serta menyusun hasil evaluasi sebagai dasar analisis di bab berikutnya. Setiap metrik digunakan sebagai acuan dalam membandingkan performa kedua algoritma yang digunakan.

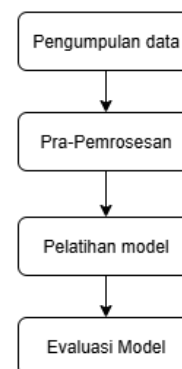
## E. Analisis dan Perbandingan

Pada tahap ini, dilakukan perbandingan performa antara model Random Forest dan Support Vector Machine berdasarkan hasil evaluasi yang diperoleh. Analisis dilakukan dengan memperhatikan kekuatan dan kelemahan masing-masing algoritma dari sisi hasil prediksi, sensitivitas terhadap data tidak seimbang, serta kompleksitas model.

Perbandingan ini dilakukan tidak hanya berdasarkan akurasi, tetapi juga berdasarkan metrik lain seperti recall, F1-score, dan ROC-AUC, yang sangat penting dalam konteks klasifikasi medis seperti prediksi stroke. Tahap ini sangat penting karena menjadi dasar dalam mengambil keputusan untuk memilih model yang paling optimal dan sesuai untuk digunakan pada sistem pendukung keputusan medis berbasis data.

## III. DESAIN, HASIL DAN PEMBAHASAN

### A. Tahapan Penelitian



Gambar 1. Tahapan Penelitian

Gambar di atas menunjukkan proses penelitian yang dilakukan dalam pembuatan dan pengecekan model prediksi risiko stroke menggunakan machine learning. Proses dimulai dengan mengumpulkan data, yaitu dengan menggunakan dataset publik yang berisi informasi tentang usia, jenis kelamin, dan kondisi kesehatan pasien. Selanjutnya, data tersebut dilakukan pra-pemrosesan, seperti mengisi data yang kosong, encoding variabel kategorikal, normalisasi fitur numerik, pembagian data latih dan uji, serta penyeimbangan kelas menggunakan metode SMOTE untuk mengatasi ketidakseimbangan distribusi antara pasien stroke dan non-stroke.

Setelah data siap, dilakukan pemodelan algoritma dengan dua jenis algoritma klasifikasi, yaitu Random Forest Classifier (RFC) dan Support Vector Machine (SVM). Kedua model tersebut dilatih menggunakan data latihan yang telah diseimbangkan, kemudian diuji dengan data uji yang telah disiapkan. Setelah pelatihan selesai, performa kedua model dinilai menggunakan beberapa metrik evaluasi, seperti akurasi, precision, recall, F1-score, dan ROC-AUC.

Setelah mendapatkan semua hasil evaluasi, dilakukan analisis dan perbandingan antara performa kedua model. Analisis ini tidak hanya memperhatikan akurasi, tetapi juga fokus pada kemampuan model dalam mendeteksi kasus stroke secara tepat. Hal ini sangat penting dalam bidang medis karena kesalahan deteksi bisa berdampak serius. Seluruh tahapan tersebut menjadi dasar untuk menentukan model mana yang lebih baik dan layak digunakan dalam sistem prediksi risiko stroke.

#### B. Rumus Evaluasi Model

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

$$\text{Precision} = \frac{TP}{TP + TF}$$

$$\text{Recall} = \frac{TP}{(TP + FN)}$$

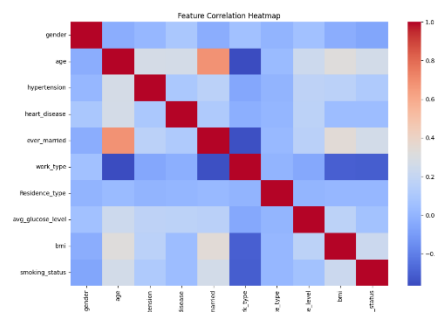
$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Rumus 1 Accuracy, Precision, Recall, dan F1-Score

Keterangan :

- TP (True Positif) : Jumlah kasus stroke yang berhasil diklasifikasikan dengan benar.
- TN (True Negative) : Jumlah data non-stroke yang diklasifikasikan dengan benar.
- FP (False Positif) : Jumlah data non-stroke yang salah diklasifikasikan sebagai stroke.
- FN (False Negative) : Jumlah data stroke yang tidak terdeteksi atau diklasifikasikan sebagai non-stroke.

#### C. Analisis Korelasi Fitur

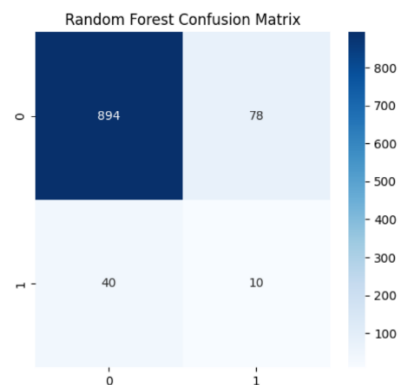


Gambar 2. Heatmap Korelasi Fitur

Gambar diatas menampilkan heatmap yang menunjukkan hubungan antar fitur dalam dataset prediksi stroke. Warna merah menggambarkan hubungan positif yang kuat, sementara warna biru menunjukkan hubungan negatif. Tingkat hubungan tersebut berkisar dari -1 hingga 1.

Dari visualisasi ini terlihat bahwa sebagian besar fitur memiliki korelasi rendah satu sama lain, yang menunjukkan tidak adanya multikolinieritas signifikan dalam data. Korelasi positif cukup terlihat antara variabel age dan hypertension, serta age dan ever\_married. Hal ini mendukung bahwa usia merupakan salah satu faktor dominan yang berkontribusi terhadap risiko stroke.

#### D. Confusion Matrix Random Forest Classifier



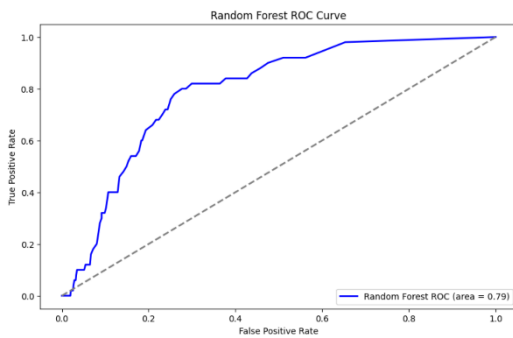
Gambar 3. Confusion Matrix Random Forest

Confusion matrix yang dihasilkan dari prediksi model Random Forest terhadap data uji, diketahui bahwa model berhasil mengklasifikasikan 894 data non-stroke secara benar (True Negative), namun juga menghasilkan 78 kesalahan klasifikasi dengan mengidentifikasi data non-stroke sebagai stroke (False Positive). Dari 50 data pasien stroke, hanya 10 kasus yang terdeteksi dengan benar (True Positive), sementara 40 sisanya tidak berhasil dikenali dan diklasifikasikan sebagai non-stroke (False Negative).

Interpretasi dari hasil tersebut dapat disimpulkan bahwa meskipun model memiliki tingkat akurasi yang tinggi secara keseluruhan, kemampuannya dalam mendeteksi kelas minoritas yaitu pasien stroke masih rendah. Jumlah false negative yang tinggi mencerminkan sensitivitas model yang lemah terhadap kasus stroke, yang

ditunjukkan oleh nilai recall sebesar 20%. Kondisi ini menjadi kekurangan yang signifikan, terutama dalam aplikasi medis yang sangat bergantung pada akurasi dalam mendeteksi kasus positif.

E. ROC Curve Random Forest Classifier

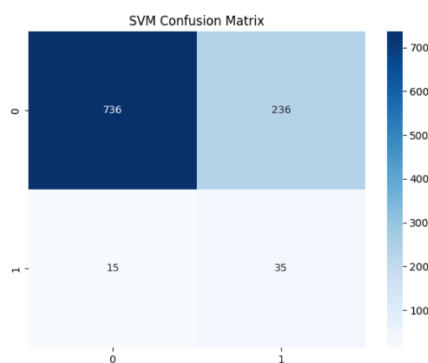


Gambar 4. ROC Curve Random Forest

ROC (Receiver Operating Characteristic) curve menunjukkan hubungan antara True Positive Rate (TPR) dan False Positive Rate (FPR) pada berbagai nilai ambang (threshold). Semakin dekat kurva tersebut ke sudut kiri atas grafik, maka semakin baik kemampuan model dalam membedakan antara kelas positif dan negatif.

Nilai Area Under the Curve (AUC) yang didapat adalah 0,79, artinya model memiliki kemampuan dalam menunjukkan mengklasifikasikan data secara umum. AUC yang mendekati nilai 1 mengindikasikan bahwa model mampu membedakan secara akurat antara pasien yang mengalami stroke dan yang tidak. Namun demikian, meskipun nilai AUC tersebut termasuk dalam kategori baik, rendahnya nilai recall dan precision menunjukkan bahwa performa model masih belum optimal, terutama dalam konteks prediksi medis yang memerlukan deteksi kasus positif dengan sensitivitas tinggi.

F. Confusion Matrix Support Vector Machine (SVM)

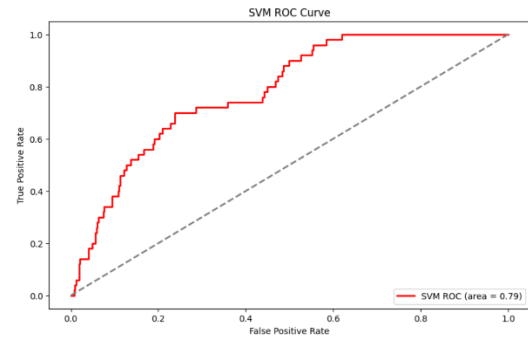


Gambar 5. Confusion Matrix SVM

Confusion matrix dari model Support Vector Machine (SVM) menunjukkan bahwa sebanyak 736 data non-stroke berhasil diklasifikasikan dengan benar (True Negative), sementara 236 data non-stroke salah diklasifikasikan sebagai stroke (False Positive). Dari 50 data pasien stroke, model ini mampu mengidentifikasi 35 kasus secara tepat (True Positive), dan hanya 15 kasus yang salah diklasifikasikan sebagai non-stroke (False Negative).

Hasil ini menunjukkan bahwa SVM memiliki kemampuan deteksi kasus stroke yang jauh lebih baik dibandingkan Random Forest, yang terlihat dari jumlah true positive yang tinggi dan nilai recall sebesar 70%. Meskipun precision-nya rendah akibat banyaknya false positive, keunggulan utama SVM adalah pada kemampuannya mengidentifikasi pasien yang berisiko stroke secara lebih sensitif, yang sangat penting dalam konteks medis.

G. ROC Curve Support Vector Machine (SVM)



Gambar 6. ROC Curve SVM

Nilai Area Under the Curve (AUC) yang diperoleh untuk model Support Vector Machine (SVM) adalah sekitar 0,785, yang menunjukkan bahwa model memiliki kemampuan yang baik dalam membedakan antara pasien yang mengalami stroke dan yang tidak. Meskipun nilainya hampir setara dengan model Random Forest, kombinasi antara AUC tersebut dan nilai recall yang tinggi menjadikan SVM lebih tepat diterapkan dalam sistem deteksi dini stroke, karena mampu mengidentifikasi kasus positif dengan sensitivitas yang lebih baik.

H. Perbandingan Model RFC dan SVM

Table 1 Perbandingan Performa Model

Metrik Evaluasi	Model Algoritma	
	Random Forest Classifier (RFC)	Support Vector Machine (SVM)
Precision	0.10	0.13
Recall	0.20	0.70
F1-Score	0.14	0.21
Accuracy	0.88	0.75

Gambar diatas menampilkan tabel perbandingan performa model antara algoritma Random Forest Classifier (RFC) dan Support Vector Machine (SVM) berdasarkan empat metrik utama: precision, recall, F1-score, dan accuracy.

- Nilai precision dari kedua model tergolong rendah, dengan SVM menunjukkan hasil yang sedikit lebih unggul dibandingkan RFC. Hal ini mencerminkan bahwa kedua model masih menghasilkan tingkat false positive yang cukup tinggi, yaitu banyak data non-stroke yang diklasifikasikan sebagai stroke.
- Recall atau sensitivitas menjadi pembeda paling mencolok antara kedua model. SVM mencatat nilai recall mendekati 0,7 (70%), sedangkan RFC hanya

mencapai sekitar 0,2 (20%). Artinya, model SVM lebih efektif dalam mengenali pasien stroke, yang merupakan aspek penting dalam sistem prediksi medis.

- F1-score, sebagai keseimbangan antara precision dan recall, juga menunjukkan keunggulan SVM. Nilai F1 SVM lebih tinggi dari RFC, yang berarti SVM memiliki kinerja yang lebih seimbang dalam memprediksi kelas positif.
- Accuracy model RFC lebih tinggi dibandingkan SVM. Namun, hal ini dapat disebabkan oleh dominasi data non-stroke dalam dataset, yang membuat model tampak lebih akurat meskipun kurang sensitif terhadap deteksi kasus stroke.

#### IV. KESIMPULAN

Penelitian ini bertujuan untuk mengevaluasi dan membandingkan performa dua algoritma klasifikasi, yaitu Random Forest Classifier (RFC) dan Support Vector Machine (SVM), dalam memprediksi risiko stroke menggunakan data pasien dari dataset publik. Proses penelitian mencakup berbagai tahapan pra-pemrosesan seperti normalisasi, encoding, dan penyeimbangan data menggunakan teknik SMOTE, kedua model dilatih dan diuji secara sistematis. Hasil evaluasi menunjukkan bahwa meskipun model RFC menghasilkan akurasi yang lebih tinggi secara keseluruhan, kemampuannya dalam mengenali kasus stroke masih rendah, terlihat dari nilai recall yang hanya mencapai 20%. Sebaliknya, SVM menunjukkan performa yang lebih konsisten dalam mendeteksi pasien stroke, dengan nilai recall mencapai 70%, meskipun akurasinya sedikit lebih rendah dibandingkan RFC.

Dari hasil perbandingan ini, dapat disimpulkan bahwa tujuan utama penelitian dalam mengevaluasi dan menentukan algoritma yang lebih efektif dalam prediksi risiko stroke telah tercapai. Model SVM dinilai lebih cocok untuk diterapkan dalam sistem prediksi risiko stroke, karena memiliki sensitivitas yang lebih tinggi terhadap kelas positif, suatu hal yang sangat penting dalam konteks medis. Hasil ini menunjukkan bahwa algoritma dengan recall tinggi lebih direkomendasikan untuk diimplementasikan dalam sistem deteksi dini penyakit kritis. Selain itu, hasil penelitian ini dapat menjadi landasan awal untuk pengembangan sistem pendukung keputusan medis berbasis machine learning, terutama untuk kasus klasifikasi penyakit dengan distribusi data yang tidak seimbang.

#### V. REFERENSI

- [1] A. P. Siregar, D. P. Purba, J. P. Pasaribu, and K. R. Bakara, "Implementasi Algoritma Random Forest Dalam Klasifikasi Diagnosis Penyakit Stroke," *J. Penelit. Rumpun Ilmu Tek.*, vol. 2, no. 4, pp. 155–164, 2023, doi: <https://doi.org/10.55606/juprit.v2i4.3039>.
- [2] D. R. Nurqotimah, A. N. Khudori, and R. S. Pradini, "Implementasi Algoritma Support Vector Machine (SVM) Untuk Klasifikasi Penyakit Stroke," *J. Appl. Comput. Sci. Technol.*, vol. 5, no. 2, p. press, 2024, doi: <https://doi.org/10.52158/jacost.v5i2.817>.
- [3] Y. A. Utama and S. S. Nainggolan, "Faktor Resiko yang Mempengaruhi Kejadian Stroke: Sebuah Tinjauan Sistematis," *J. Ilm. Univ. Batanghari Jambi*, vol. 22, no. 1, p. 549, 2022, doi: <http://dx.doi.org/10.33087/jiubj.v22i1.1950>.
- [4] A. Setiawan, R. F. Waleska, M. A. Purnama, Rahmaddeni, and L. Efrizoni, "Komparasi ALgoritma K-Nearest Neighbor (K-NN), Support Vector Machine (SVM), dan Decision Tree Dalam Klasifikasi Penyakit Stroke," *J. Inform. Rekayasa Elektron.*, vol. 7, no. 1, p. 3, 2024, doi: <https://doi.org/10.36595/jire.v7i1.1161>.
- [5] C. L. Bhuana and H. W. Nugroho, "Perbandingan Decision Tree C4.5 dan Support Vector Machine (SVM) Dalam Klasifikasi Penderita Stroke Berbasis PSO," *J. SIMADA (Sistem Inf. dan Manaj. Basis Data)*, vol. 6, no. 1, pp. 24–34, 2023, doi: 10.30873/simada.v6i1.3429.
- [6] F. Fadmadika, H. H. Handayani, T. Al Mudzakir, and J. Indra, "Pengaruh Smote Terhadap Performa Algoritma Random Forest Dan Algoritma Gradient Boosting Dalam Memprediksi Penyakit Stroke," *J. Tek. Inf. dan Komput.*, vol. 7, no. 2, p. 837, 2024, doi: <https://doi.org/10.37600/tekinkom.v7i2.1575>.
- [7] R. Wahyudi, K. I. Manik, M. Alfin, J. B. Henrydunan, M. H. Al Majid, and K. Saputra, "Klasifikasi Penyakit Migrain Menggunakan Metode Support Vector Machine," *J. Manaj. Inform. Sist. Inf.*, vol. 8, no. 2, pp. 176–187, 2025, doi: <https://doi.org/10.36595/misi.v8i2.1587>.
- [8] Alwaliyanto, S. K. Gusti, I. Afrianty, and F. Syafria, "Penerapan Metode ADASYN Dalam Mengatasi Imbalanced Data Untuk Klasifikasi Penyakit Stroke Menggunakan Support Vector Machine," *Media Online*, vol. 5, no. 4, pp. 532–541, 2025, doi: <https://doi.org/10.47065/bulletincsr.v5i4.612>.
- [9] A. R. Dana, R. V. Kristananda, M. B. S. Wibowo, and D. A. Prasetya, "Perbandingan Algoritma Decision Tree dan Random Forest dengan Hyperparameter Tuning dalam Mendeteksi Penyakit Stroke," *Semin. Nas. Inform. Bela Negara*, vol. 4, pp. 66–75, 2024, [Online]. Available: <https://santika.upnjatim.ac.id/submissions/index.php/santika/article/view/349>
- [10] A. Ristyawan, A. Nugroho, and T. K. Amarya, "Optimasi Preprocessing Model Random Forest untuk Prediksi Stroke," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 12, no. 1, 2025, doi: <https://doi.org/10.35957/jatisi.v12i1.9587>.
- [11] J. D. Muthohhar and A. Prihanto, "Analisis Perbandingan Algoritma Klasifikasi untuk

Penyakit Jantung,” *J. Informatics Comput. Sci.*,  
vol. 04, pp. 298–304, 2023, doi:  
10.26740/jinacs.v4n03.p298-304.