[1]Makalah Project Computational Biology

Analisa Kejadian Stroke Menggunakan Metode Prediction

Bidang Kegiatan :
PKM - KC (Karya Cipta)

Kelompok 2

Disusun oleh:

Calvin Lim	Ketua	2502015762	Binusian 2025
Michael Vic Chow	Anggota	2502037454	Binusian 2025
Kenichi Halim	Anggota	2502039163	Binusian 2025

UNIVERSITAS BINA NUSANTARA KEMANGGISAN
2023

1

Daftar Isi

Bab 1 Pendahuluan

1.1 Latar Belakang

Stroke merupakan penyakit serius yang terjadi ketika gumpalan darah menghambat aliran darah dan oksigen ke otak, yang mengakibatkan kerusakan sel dan situasi yang berpotensi mengancam jiwa[1]. Stroke berkontribusi besar atas kematian dan kelumpuhan di seluruh dunia dengan memasuki peringkat nomor 2 dan peringkat nomor 3 dalam penyebab terjadinya disability manusia pada tahun 2019 menurut WHO, dan bertanggung jawab atas perkiraan 6,2 juta atau 11% dari total kematian secara global yang berkembang setiap tahun[2]. Dampak stroke dapat bervariasi dari jangka pendek hingga panjang dan dapat menyebabkan kelumpuhan otot yang mempengaruhi aktivitas sehari-hari, seperti kemampuan bicara, menggerakkan tubuh, dan lain lain [3]. Luasnya dampak ini dipengaruhi oleh seberapa banyak bagian otak yang terpengaruh dan cepatnya pengobatan yang diberikan. Oleh karena itu, deteksi dan intervensi dini yang tepat waktu sangat penting untuk meminimalisir dampak stroke[4].

Penelitian ini ditujukan untuk mengidentifikasi faktor risiko yang dapat mempengaruhi kemungkinan seseorang mengalami stroke di masa yang mendatang. Kami akan menggunakan empat metode prediksi: Linear SVC, KNeighbors, Logistic Regression, dan Linear Regression. Metode-metode ini telah digunakan dalam berbagai penelitian untuk klasifikasi dan prediksi. Misalnya, Linear SVC telah digunakan dalam penelitian untuk kategorisasi efektif, penyaringan berita, personalisasi, dan pengarahan informasi. Logistic Regression juga digunakan dalam penelitian untuk masalah klasifikasi dan juga disebut Linear Regression[5].

Dengan memahami pola kejadian stroke dan faktor-faktor risiko yang berkontribusi, penelitian ini berharap dapat memberikan wawasan yang berharga untuk pencegahan dan pengobatan stroke. Selain itu, penelitian ini juga diharapkan untuk berkontribusi pada bidang biocomputation yang menyatukan biologi dan juga prediksi model matematika yang dilakukan oleh mesin melalui prediksi dari empat model yang digunakan, dan memahami perbandingan prediksi stroke antara empat model yang digunakan.

1.2 Rumusan Masalah

1. Bagaimana menganalisa dataset pasien yang dikumpulkan untuk memprediksi kejadian stroke?
2. Bagaimana memilih dan mengolah fitur-fitur yang relevan dalam data untuk meningkatkan akurasi prediksi?
3. Bagaimana mengidentifikasi pola yang berkaitan dengan kejadian stroke?
4. Apa perbandingan antara ke empat model yang digunakan?

1.3 Tujuan Masalah

1. Mengetahui pola yang berkaitan dengan kejadian kasus Stroke
2. Meminimalkan dampak terjadinya stroke dengan melakukan pencegahan
3. Memberikan kontribusi pada bidang medis melalui prediksi dari keempat model
4. Mengetahui perbandingan prediksi stroke antara keempat model yang digunakan

Bab 2 Methodology

2.1 Dataset

Pengumpulan dataset dilakukan terlebih dahulu agar dapat digunakan di penelitian ini. Dataset yang kami gunakan berasal dari Website Kaggle dengan judul dataset Stroke Prediction Dataset. Tujuan dari dataset ini adalah melakukan analisa prediksi apakah seorang pasien menderita penyakit stroke melalui karakteristik individu yang diberikan. Dataset ini terdiri atas 5110 data pasien yang memiliki 12 variable. 12 variable ini merupakan data informasi pribadi masing-masing pasien. 12 variabel berikut dapat diperhatikan pada tabel 1 dibawah ini.

TABLE I

Variabel yang terdapat pada Dataset

Variable	**Description**	**Value**
ID	ID unique pasien	Angka
Gender	gender pasien	0 = Female 1 = Male
age	umur pasien sekarang	Angka
Hypertension	apakah pasien mengalami hipertensi?	0 = no 1 = yes
heart_disease	apakah pasien mengalami penyakit hati?	0 = no 1 = yes
ever_married	status nikah pasien	0 = no 1 = yes
work_type	status kerja pasien	0 = Govt_job 1 = Never_worked 2 = Private 3 = Self-employed 4 = Children
Residence_type	tempat tinggal	0 = Rural 1 = Urban
avg_glucose_level	rata-rata level glukosa darah pasien	Angka
bmi	body mass index	Angka
Obesity	tingkat obesitas pasien	0 = Underweight 1 = Healthy

		2 = OverWeight 3 = Obesitas
smoking_status	Status pasien mengenai merokok	0 = Unknown 1 = Formerly smoked 2 = Never smoked 3 = Smokes
stroke	apakah pasien menghadapi stroke?	0 = no 1 = yes

2.2 Data Preprocessing

Data mentah dapat mengurangi hasil akhir kualitas prediksi, baik karena data yang kosong dan data rusak. Maka dari itu, agar data yang digunakan tidak menyebabkan kecacatan dalam model diperlukan melakukan data preprocessing. Tahap-tahap yang dilakukan dibagi ke dalam beberapa segmen, yaitu:

2.2.1 Missing values treatment

Awal mula preprocessing dimulai dengan melibatkan identifikasi dan melakukan perawatan dengan cara penggantian nilai, ada beberapa metode untuk mengganti nilai yang kosong, seperti jika parameternya berupa numeric maka nilai yang hilang dapat diganti dengan nilai rata-rata (mean), nilai tengah (median) parameter, regression imputation, dan berbagai cara lainnya. Kemudian untuk missing values pada categorical parameter dapat diisi dengan modus, membuat kategori baru, dan cara lainnya.

2.2.2 Feature selection

Feature selection merupakan tahap untuk mengurangi parameters yang akan digunakan pada pembelajaran model dengan cara memilih features yang relevan dan informatif yang ada pada dataset, sehingga data yang tidak relevan tidak menginterupsi dan menyebabkan bias pada pembelajaran model.

2.2.3 Encoding Categorical Data

Data categorical biasanya merupakan text dan hal ini menyebabkan tidak praktisnya pembelajaran model. Dengan melakukan encoding terhadap categorical data, maka data yang dimiliki menjadi lebih mudah untuk dipahami dalam pembelajaran model. Tahap ini dapat dilakukan dengan metode label encoding, binary encoding, ordinal encoding, dan metode lainnya.

2.3 Machine Learning Models

2.3.1 Linear Support Vector Classification (SVC)

Linear Support Classification (SVC) merupakan metode klasifikasi biner untuk memisahkan antar kelas pasien yang ada dengan cara mencari suatu titik untuk data satu dimensi, garis untuk data dua dimensi, ataupun bidang untuk data dengan banyak dimensi dengan hasil titik, garis, ataupun bidang yang paling optimal yang dapat memisahkan

klasifikasi antar kelas pasien, sehingga titik data antar kelas pasien terpisahkan oleh titik, garis, ataupun bidang yang ditentukan.

2.3.2 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) adalah metode yang digunakan untuk klasifikasi dan regresi. Dalam konteks ini, KNN digunakan untuk klasifikasi data pasien berdasarkan kemiripan fitur mereka dengan pasien lain dalam dataset. KNN bekerja dengan mengidentifikasi K titik data terdekat dan mengklasifikasikan titik data baru berdasarkan mayoritas kelas dari titik data terdekat tersebut. Jarak ini biasanya dihitung menggunakan rumus jarak Euclidean:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

di mana d adalah jarak, dan (x1,y1) dan (x2,y2) adalah koordinat dua titik dalam ruang dua dimensi.

2.3.3 Logistic Regression

Logistic Regression adalah metode yang digunakan untuk klasifikasi biner. Dalam konteks ini, Logistic Regression digunakan untuk memprediksi apakah pasien berisiko mengalami stroke berdasarkan fitur-fitur mereka. Logistic Regression bekerja dengan menggunakan fungsi logistik untuk memprediksi probabilitas suatu peristiwa.. Rumus dasarnya adalah:

$$p = \frac{1}{1 + e^{-(b_0 + b_1 x)}}$$

di mana p adalah probabilitas hasil positif, b0 dan b1 adalah koefisien regresi, x adalah variabel independen, dan e adalah basis logaritma natural.

2.3.4.Gaussian NB

Gaussian Naive Bayes (NB) adalah metode yang digunakan untuk klasifikasi. Dalam konteks ini, Gaussian NB digunakan untuk memprediksi apakah pasien berisiko mengalami stroke berdasarkan fitur-fitur mereka. Gaussian NB bekerja dengan mengasumsikan bahwa fitur-fitur mengikuti distribusi normal atau Gaussian.. Rumus dasarnya adalah:

$$P(y|X) = P(X|y) * P(y) / P(X)$$

di mana P(y|X) adalah probabilitas posterior, P(X|y) adalah probabilitas likelihood, P(y) adalah probabilitas prior, dan P(X) adalah probabilitas marginal.

Bab 3 Experiment and Result

3.1 Eksplorasi Dataset (EDA)

Analisis data diawali dengan memperoleh pemahaman mendasar dari dataset yang digunakan, baik itu jumlah baris maupun tipe-tipe parameter.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5110 entries, 0 to 5109
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   id                 5110 non-null   int64
 1   gender             5110 non-null   object
 2   age                5110 non-null   float64
 3   hypertension       5110 non-null   int64
 4   heart_disease      5110 non-null   int64
 5   ever_married       5110 non-null   object
 6   work_type          5110 non-null   object
 7   Residence_type     5110 non-null   object
 8   avg_glucose_level  5110 non-null   float64
 9   bmi                4909 non-null   float64
 10  smoking_status     5110 non-null   object
 11  stroke             5110 non-null   int64
dtypes: float64(3), int64(4), object(5)
memory usage: 479.2+ KB
```

Gambar 1. Info Dataset

Pada Gambar 2 diperlihatkan parameter terdiri atas 5110 data pasien, dua belas parameter yang terdiri atas empat tipe integer, tiga tipe float, dan lima tipe object dan besar memori yang digunakan sebesar 492.2 KB.

3.2 Data Preprocessing

3.2.1 Missing Value Treatment

Pada gambar 1 dapat diketahui bahwa terdapat value yang missing pada salah satu parameter tersebut yaitu parameter bmi dengan jumlah data 4909. Karena tipe parameter adalah angka maka kita akan menggantikan missing value dengan nilai mean parameter bmi supaya tidak mempengaruhi kualitas data.

3.2.2 Feature Selection

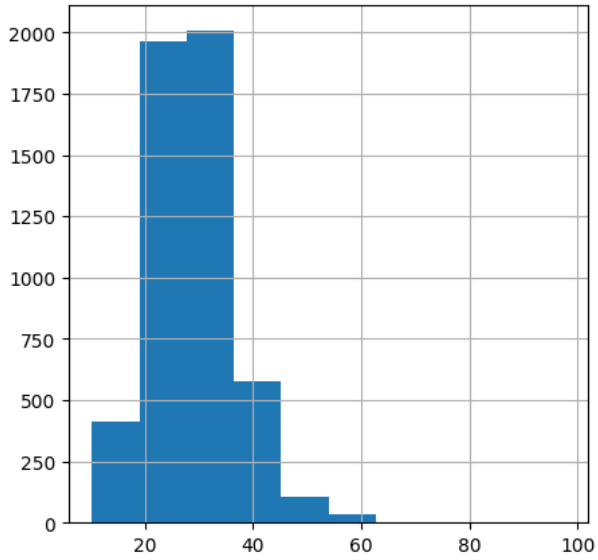
Pada Gambar 2 terdapat data sampah “Other” pada parameter gender, jadi data tersebut kami tangani dengan cara menggantikan kata “Other” dengan kategori value yang sering muncul (Modus)

```
Column name: gender

Unique values:
Female    2994
Male      2115
Other        1
Name: gender, dtype: int64
```

Gambar 2. Data Unique Gender

Pada Gambar 3 jika kita melihat visualisasi grafik parameter bmi kita dapat melihat bahwa data yang paling sering muncul adalah data antara dua puluh hingga empat puluh sedangkan data lebih dari empat puluh lebih sedikit, data yang sedikit ini dapat mempengaruhi kualitas data untuk menanganinya kita membuat kategori parameter baru “obesity” yang dihasilkan dari parameter bmi dengan kondisi yaitu, value 0 jika bmi < 18.5, value 1 jika bmi >= 18.5 dan bmi < 25, value 2 jika bmi >= 25 dan bmi < 30, dan value 3 jika bmi >=30.



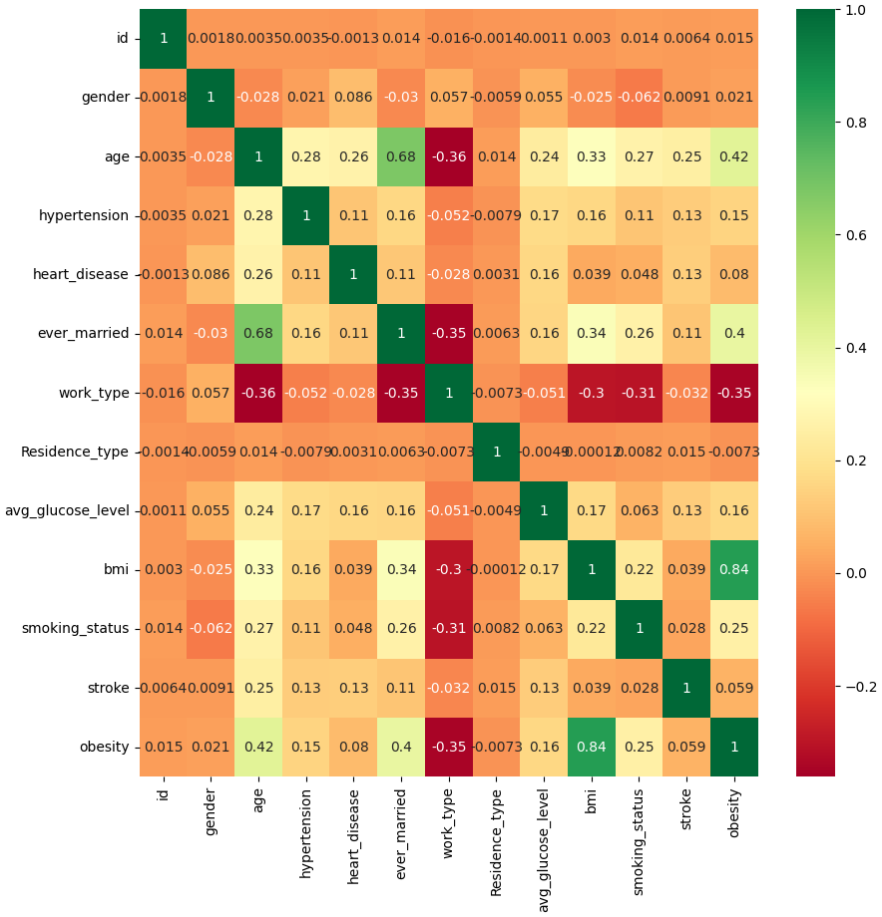
Gambar 3. Graphic bmi

3.2.3 Encoding

Mengubah semua parameter yang bertipe object menjadi category angka. Ini bertujuan supaya kita dapat melakukan visualisasi graphic dan model training yang hanya bisa diproses dalam bentuk angka

3.2.4 Visualization After Preprocessing

Pada Gambar 4. Plot heatmap Kita dapat analisa bahwa coor untuk parameter "id","gender", "bmi","work_type","gender", "smoking_status","Residence_type" sangatlah rendah, maka dari itu diperlukan untuk membuang data untuk meningkatkan kualitas data



Gambar 4. Coor HeatMap

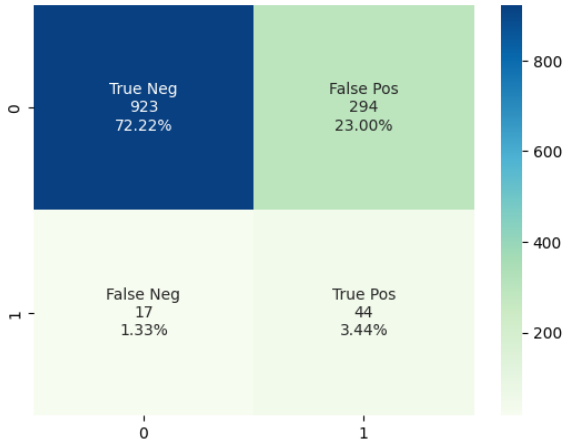
3.3 Data Training

Melakukan pemisahan data independent dan dependent dimana dependent adalah parameter stroke. Pembagian Jumlah dataset 5100 dibagi secara acak dengan ratio 3 : 1, dimana 75% diberikan pada data training dan 25% pada data testing dari data sebenarnya. Distribusi value akhir pada parameter stroke tidaklah seimbang dimana value tidak stroke muncul sebesar 95% dan stroke muncul sebesar 5%, dan karena keseimbangan antar data sangat penting kita akan menggunakan metode SMOTE untuk menyeimbangkan value akhir menjadi sama rata.

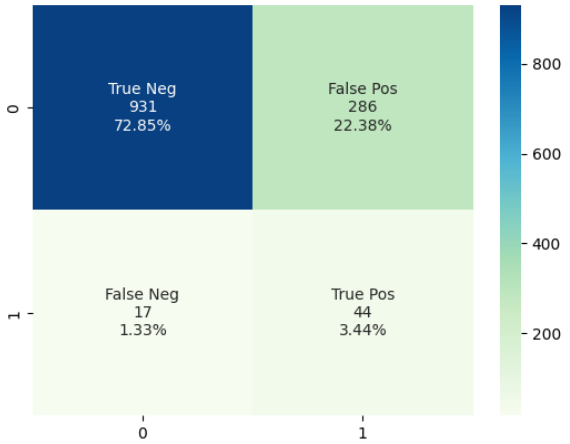
3.4 Evaluation

Gambar 5, 6, 7 dan 8 menunjukan performa keempat model ML, dalam bentuk confusion matrix, warna dengan korelasi besar adalah warna biru sedangkan korelasi terkecil dengan warna hijau putih. True Neg adalah data bukan penderita Stroke(Neg) yang diprediksi benar(True) / tidak derita Stroke. True Pos adalah data penderita Stroke(Pos) yang diprediksi benar(True) / menderita stroke. False Pos adalah data penderita Stroke(Pos) yang diprediksi salah(False) / tidak derita Stroke. False Neg adalah data bukan penderita Stroke(Neg) yang diprediksi salah(False) / derita stroke.

False Neg yang tingkat persentase tinggi sangatlah berbahaya dalam kasus prediksi Stroke. Jika pasien menderita Stroke tetapi diprediksi tidak menderita (FN), maka pasien akan mengetahui keadaan sebenarnya dengan sangat lambat dan pasien tersebut tidak segera melakukan perawatan medis untuk Stroke. Untuk prediksi Stroke, penting untuk memprioritaskan dalam meminimalkan False Neg sebanyak mungkin.



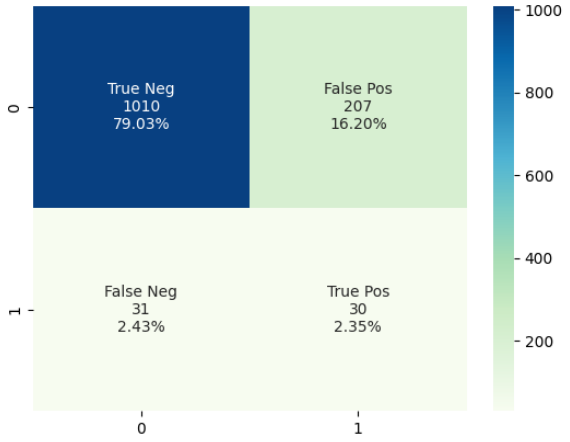
Gambar 5. Confusion Matrix Linear SVC



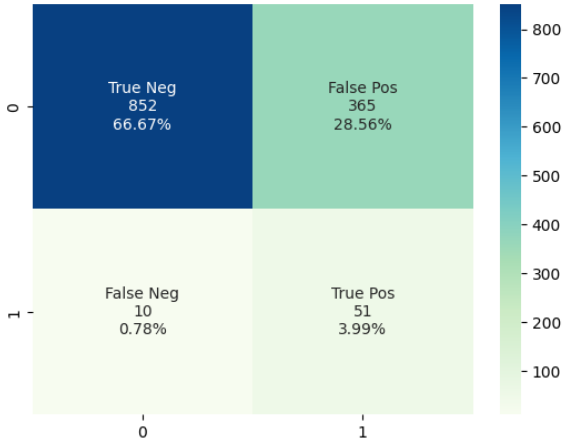
Gambar 6. Confusion Matrix Logistic Regression

Linear SVC dapat diketahui bahwa terdapat 967 prediksi benar dengan akurasi 76%. Namun 311 data salah diprediksi, dengan 17 sebesar 1% kasus Stroke dengan prediksi sebagai non-stroke dan 294 sebesar 23% kasus non-stroke dengan prediksi sebagai stroke.

Logistic Regression memiliki model yang lebih bagus daripada Linear SVC meskipun sangat sedikit, dengan akurasi True Neg adalah 72.85% lebih tinggi 0.63% dari TN SVC, dan akurasi False Pos sebesar 22.38% yang lebih kecil 0.62% dari FP Linear SVC.



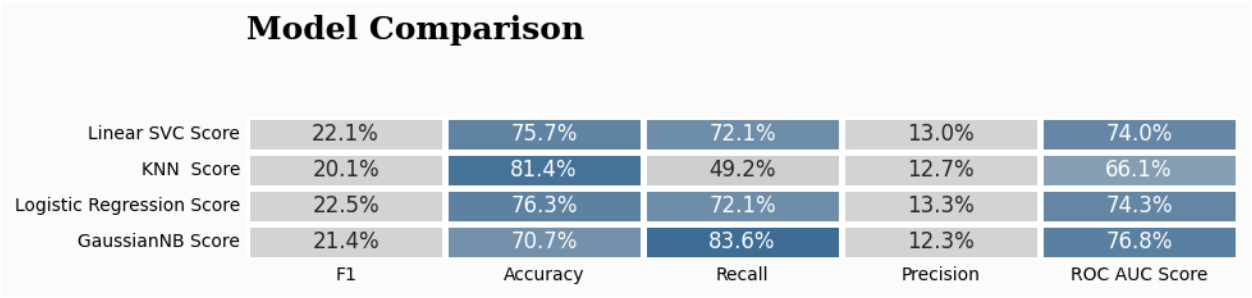
Gambar 7. Confusion Matrix K-Nearest Neighbour (KNN)



Gambar 8. Confusion Matrix Gaussian Naive Bayes

K-Nearest Neighbour (KNN) menghasilkan akurasi tertinggi sebesar 81% tapi sayangnya dia menghasilkan prediksi akurasi tertinggi pada False Neg sebesar 2% dimana telah terjadi 31 kasus dari total 1278 kejadian.

Gaussian Naive Bayes merupakan model yang menghasilkan akurasi paling kecil diantara 4 model dengan besar 71% dan juga menghasilkan False Pos terbesar dengan besar 29% sebanyak 365 kasus, tetapi Naive Bayes juga merupakan model yang menghasilkan akurasi False Neg terkecil sebesar 0.8% lebih kecil dari 1%, klasifikasi FN yang paling krusial dalam memprediksi penderita Stroke



Gambar 9. Model Comparison

Berdasarkan Gambar 9, kita dapat simpulkan bahwa model KNN merupakan model terbaik dari model lainnya dalam menunjukan tingkat prediksi yang kuat secara keseluruhan dengan akurasi 81%, namun KNN tidak menunjukan tingkat prediksi yang kuat dalam hal score seperti score f1 20%, Precision 13%, ROC AUC 66%, dan Recall dengan score 49% yang sangat kecil daripada model lainnya. Model Logistic Regression mempunyai kinerja yang sedikit lebih bagus daripada model Linear SVC dalam semua score f1, accuracy, Recall, Precision dan ROC AUC.

Model Gaussian Naive Bayes mengungguli dalam kinerja kerja pada score Recall dan ROC AUC dengan tingkat prediksi 84% dan 77%, model ini memiliki akurasi yang cukup bagus 71% meskipun tingkatnya paling kecil daripada model lainnya serta tingkat score f1 sebesar 21% yang lebih besar daripada score KNN. Model Naive Bayes merupakan model yang paling cocok dalam pendeteksian Stroke karena tingkat akurasi False Neg yang sangat rendah dapat mengurangi dampak dan tingkat fatalitas manusia akibat dari deteksi stroke yang salah.

Bab 4 Conclusion

Stroke merupakan sebuah ancaman dalam kehidupan manusia dan harus segera dicegah dan dirawati segera mungkin setelah dideteksi. Dengan perkembangan dunia teknologi yang cepat, teknologi AI/ML dapat membantu dokter dan ilmuwan dalam menemukan parameter paling relevan dalam terjadinya stroke. Pada analisa ini kami mengusulkan untuk melakukan selection feature sebelum melakukan perbandingan antara keempat model KNN, Linear SVC, Logistic Regression dan Gaussian Naive Bayes untuk prediksi terjadinya Stroke.

Performa dalam mengevaluasi klasifikasi menggunakan f1, accuracy, Recall, Precision dan AUC sangat penting untuk interpretasi model dan demonstrasi performa klasifikasi.

Confusion matrix juga perlu diperhatikan untuk mengurangi pendeteksi yang salah. Model Naive Bayes mengungguli metode lainnya, dengan Recall sebesar 84%, ROC AUC sebesar 77% dan kemungkinan terjadi False Neg sebesar 0.8%. Karena itu, metode Naive Bayes adalah metode paling sesuai dalam mengidentifikasi dan mengamankan pasien dari resiko tinggi terjadinya stroke dalam jangka panjang. Penelitian ini masih perlu dilanjutkan untuk menemukan model yang lebih efisien untuk menghasilkan prediksi yang tinggi dan tingkat False Neg seminimal mungkin.

Reference

1. Dritsas, E., & Trigka, M. (2022). Stroke Risk Prediction with Machine Learning Techniques. Sensors (Basel, Switzerland), 22(13), 4670. doi: 10.3390/s22134670
2. Feng, X., Liu, C., Guo, Q., Bai, Y., Ren, Y., Ren, B., Bai, J., & Chen, L. (2013). Research progress in rehabilitation treatment of stroke patients: A bibliometric analysis. Neural regeneration research, 8(15), 1423–1430. https://doi.org/10.3969/j.issn.1673-5374.2013.15.010
3. R. Ali, U. Qidwai and S. K. Ilyas, Use of Combination of PCA and ANFIS in Infarction Volume Growth Rate Prediction in Ischemic Stroke. 2018 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES), Sarawak, Malaysia, 2018, pp. 324-329. doi: 10.1109/IECBES.2018.8626629.
4. Kato, T., Lee, S., & Narayanan, S.S. (2009). An analysis of articulatory-acoustic data based on articulatory strokes. 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, 4493-4496. doi: 10.1109/ICASSP.2009.4960628
5. Kim, H., Kim, Y. H., Kim, S. J., & Choi, M. T. (2022). Pathological gait clustering in post-stroke patients using motion capture data. Gait & posture, 94, 210–216. doi: 10.1016/j.gaitpost.2022.03.007