Journal of Information Technology and Computer Science (INTECOMS)
Volume 7 Nomor 4, Tahun 2024
e-ISSN : 2614-1574
p-ISSN : 2621-3249

MODEL PREDIKSI RISIKO STROKE MENGGUNAKAN MACHINE LEARNING

STROKE RISK PREDICTION MODEL USING MACHINE LEARNING

Iwa Ovyawan Herlistiono[1], Sriyani Violina[2]
[1,2]Informatics Dept., Universitas Widyatama, Jalan Cikutra No. 204A, Bandung
ovyawan.herlistiono@widyatama.ac.id

ABSTRACT

According to data from the Social Security Administration (BPJS), stroke is included in the top four catastrophic diseases, which require high costs for treatment and have complications that can be life-threatening. Meanwhile, according to data from the Ministry of Health, around 80% of Indonesian people do not know the symptoms of stroke, so stroke treatment is often too late. In this research, machine learning methods were applied to predict the risk of stroke. Machine Learning has often been used in the health sector, especially to predict disease risk and classify certain diseases based on patient data. In this study, publicly available patient data was used to design a stroke risk prediction model consisting of 5000 pieces of data. This study makes an important contribution to the field of stroke prevention by introducing a risk prediction model that can help identify individuals at high risk for further management. The application of a stroke risk prediction model using machine learning is expected to increase early detection and timely intervention, as well as reduce the overall burden of stroke. In addition, this research also provides a basis for the development of more sophisticated and effective stroke risk prediction models in the future.

***Keyword** : Machine Learning, Stroke, Prediction.*

ABSTRAK

Menurut data Badan Penyelenggaraan Jaminan Sosial (BPJS), stroke termasuk dalam empat besar penyakit katastropik, yang membutuhkan biaya tinggi dalam pengobatannya dan memiliki komplikasi yang dapat mengancam jiwa. Sementara itu menurut data Kementerian Kesehatan, sekitar 80% masyarakat Indonesia tidak mengetahui gejala stroke sehingga seringkali penanganan stroke menjadi terlambat. Pada penelitian ini dilakukan penerapan metode machine learning untuk memprediksi risiko penyakit stroke. Machine Learning sudah sering digunakan di bidang kesehatan terutama untuk memprediksi risiko penyakit dan klasifikasi penyakit tertentu berdasarkan data pasien. Pada penelitian ini digunakan data pasien yang tersedia secara publik untuk merancang model prediksi risiko penyakit stroke yang terdiri dari 5000 data. Studi ini memberikan kontribusi penting dalam bidang pencegahan stroke dengan memperkenalkan model prediksi risiko yang dapat membantu identifikasi individu dengan risiko tinggi untuk pengelolaan lebih lanjut. Penerapan model prediksi risiko stroke menggunakan machine learning, diharapkan dapat meningkatkan deteksi dini dan intervensi yang tepat waktu, serta mengurangi beban stroke secara keseluruhan. Selain itu, penelitian ini juga memberikan dasar untuk pengembangan model prediksi risiko stroke yang lebih canggih dan efektif di masa depan.

Kata Kunci : Machine Learning, Stroke, Prediksi.

PENDAHULUAN

Stroke ialah penyakit kardiovaskuler yang terjadi akibat gagalnya suplai oksigen ke sel-sel otak, yang beresiko terhadap kerusakan iskemik dan dapat menyebabkan kematian. Di Indonesia terdapat sekitar 550.000 pasien baru stroke setiap tahunnya. Angka ini terbilang sangat tinggi dan menempati urutan ketiga sebagai penyebab kematian di Indonesia, setelah kardiovascular dan kanker. Dari segi usia, 72% pasien stroke berumur di atas 65 tahun, namun seiring dengan perubahan gaya hidup maka, kecenderungan pasien untuk mendapatkan stroke terjadi pada usia lebih muda.

Stroke merupakan suatu episode disfungsi sistem saraf karena adanya gangguan pada aliran darah ke otak, yang ditandai dengan adanya gejala klinis defisit neurologis. Stroke dapat dibuktikan dengan pemeriksaan imaging atau patologi, serta dapat mengakibatkan cacat permanen atau kematian jika tidak ditangani segera.

Menurut data Kementerian Kesehatan, sekitar 80% masyarakat Indonesia tidak mengetahui gejala stroke

sehingga seringkali penanganan stroke menjadi terlambat. Padahal, perawatan cepat dapat mengurangi kerusakan otak yang akan disebabkan oleh stroke. Bahkan di Indonesia, menurut data Badan Penyelenggaraan Jaminan Sosial (BPJS), stroke termasuk dalam empat besar penyakit katastropik, yang membutuhkan biaya tinggi dalam pengobatannya dan memiliki komplikasi yang dapat mengancam jiwa.

Penemuan dan pengendalian faktor risiko stroke dilakukan pada orang sehat, penderita yang sudah terdata mempunyai faktor risiko stroke atau pada keluarga penderita yang pernah mengalami serangan stroke.

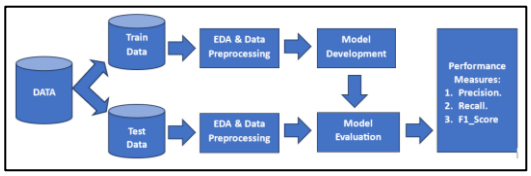
Penemuan dan pengendalian faktor risiko stroke perlu dilakukan untuk mengurangi dan mencegah keparahan dari stroke. Pengendalian risiko stroke biasanya dilakukan pada orang sehat, penderita yang sudah terdata mempunyai faktor risiko stroke atau pada keluarga pasien yang pernah mengalami serangan stroke. Jika pada seseorang terdapat faktor-faktor risiko stroke maka orang tersebut disebut sebagai *stroke prone profil.*

Deteksi dini penyakit stroke memungkinkan pasien dan dokter untuk mengambil tindakan yang tepat dan dapat lebih memahami risiko terkait. Deteksi dini merupakan suatu rangkaian kegiatan aktif untuk menemukan faktor risiko stroke.

Machine Learning sudah sering digunakan di bidang kesehatan terutama untuk memprediksi risiko penyakit dan klasifikasi penyakit tertentu berdasarkan data pasien. Pada penelitian ini digunakan data pasien yang tersedia secara publik untuk merancang model prediksi risiko penyakit stroke yang terdiri dari minimal 5000 data

METODE

Workflow Prediksi risiko penyakit stroke dimulai dengan pengumpulan data. Proses selanjutnya adalah inspeksi data menggunakan Exploratory Data Analysis (EDA) untuk persiapan proses preprocessing yang melibatkan sejumlah metode lainnya. Setelah itu data di-feed ke algoritma untuk melakukan proses training. Proses training itu akan menghasilkan model yang selanjutnya akan diuji dengan menggunakan data test. Hasilnya diukur dengan menggunakan kriteria Precision, Recall dan F1_Score. Proses-proses tersebut disarikan pada Gambar 1 berikut.



Gambar 1. Workflow Prediksi Risiko Stroke

Pengumpulan dan Pengambilan Data.

Pengumpulan data adalah tahap awal dan satu langkah terpenting dalam pengembangan model ML. Data harus relevan, berkualitas tinggi, dan memadai dari segi jumlah, untuk menghasilkan model yang dapat mempelajari pola dengan tepat dan membuat prediksi akurat.

Pedoman terpenting dari pengumpulan data adalah data tersebut harus didapatkan secara legal. Untuk lebih jelasnya bisa mengacu pada aturan dan petunjuk dari dua sumber “The European Union’s General Data Protection Regulation (GDPR)” [2] dan California Consumer Privacy Act (CCPA) [3]. Mengabaikan legalitas pengumpulan data (berarti, data didapat secara illegal), membuka peluang pada peneliti untuk mendapatkan kasus hukum.

Persyaratan berikutnya adalah relevansi data dengan riset yang akan dilakukan. Data yang akan digunakan sedapat mungkin berukuran sangat besar untuk bisa dikatakan memadai. Data juga harus mempunyai kualitas yang baik, namun hal ini bisa didapatkan dengan melakukan Data Preprocesing dengan seksama.

Berdasarkan pada pertimbangan-pertimbangan di atas, Dataset “stroke.csv” dipilih dan didapat dari public-repository

Kaggle [4]. Deskripsi lengkap dataset, memiliki 12 atribut (Tabel 1). Dataset juga memiliki ukuran yang cukup besar, 5110 baris.

Tabel 1. Deskripsi Dataset

Atribut	Keterangan/Isi Atribut
id	identifikasi unik untuk pasien.
gender	"Male", "Female" atau "Other"
age	Usia pasien
hypertension	1 jika pasien menderita hipertensi, 0 jika tidak
heart_desease	1 jika pasien menderita sakit jantung, 0 jika tidak
ever_married	"No" atau "Yes"
work_type	"children", "Govt_jov", "Never_worked", "Private" atau "Self-employed"
Residence_type	"Rural" atau "Urban"
avg_glucose_level	rata-rata tingkat gula dalam darah
bmi	body mass index.
smoking_status	"formerly smoked", "never smoked", "smokes" atau "Unknown"
stroke	1 jika pasien terserang stroke, 0 jika tidak.

Dataset tersebut dipilih karena relevan dengan tujuan penelitian dan dalam deskripsinya disebutkan, bahwa dataset tersebut berisi sejumlah informasi tentang pasien yang dapat digunakan untuk memprediksi apakah seorang pasien memiliki resiko tinggi terserang stroke.

Prediksi dibuat berdasarkan pada input berupa parameter: jenis kelamin, usia, riwayat hipertensi, riwayat penyakit jantung, status pernikahan, jenis pekerjaan, wilayah tempat tinggal, kadar gula darah, body mass index, dan apakah pasien merokok.

Setelah dataset didapatkan, sangat krusial untuk melakukan *split-dataset* sebelum melakukan proses apapun (terhadap dataset). *Split-Dataset* sebelum proses apapun ini, dilakukan untuk menghindari terjadinya *data-leaking*. Suatu kondisi saat algoritma mulai membangun model, yang seharusnya hanya berdasarkan pada Training-Data, terpengaruh oleh Test-data akibat langkah-langkah preprocessing. Misal, akibat penggunaan teknik interpolasi median atau mean, dan atau proses standarisasi atau normalisasi.

Pada penelitian ini dataset dibagi secara fisik, menjadi dua buah file .csv: stroke_train_80.csv.txt, dan stroke_test_80.csv.txt.

Exploratory Data Analysis (EDA)

EDA adalah proses eksplorasi pada dataset untuk mendapatkan karakteristik dasar dari data. Secara umum ada empat langkah dalam EDA:

1. Analisis distribusi kolom. Distribusi setiap kolom dalam dataset dieksplorasi dan dihitung secara statistik. Hasilnya adalah tipe dari tiap-tiap kolom dan statistiknya serta gambaran dari kolom tersebut berupa hasil plotting.
2. Analisis korelasi. Korelasi antar kolom dalam dataset dihitung menggunakan sejumlah matriks korelasi. Hasilnya adalah gambaran dari korelasi dengan berbagai pengukuran.
3. Analisis *missing-values*. dilakukan analisis pada tiap-tiap kolom untuk mendapatkan kolom mana saja yang berisi missing-value dan berapa jumlahnya serta pengaruhnya terhadap dataset. Hasilnya digambarkan sebagai sebuah plot.
4. Analisis perbedaan kolom (column differences). Perbedaan distribusi dan statistik pada kolom tertentu dari dua atau lebih dataset dihitung, contoh: dilakukan pada Training-data dan Test-Data setelah dataset di-split. Proses ini akan mendeteksi tipe tiap-tiap kolom dan statistiknya. [1]

Data Preprocesing.

Pada saat mendapatkan data, konsensus umum adalah selalu beranggapan bahwa data tidak sempurna. EDA dan Data preprocessing menjadi instrumen penting dalam melakukan perbaikan data.

1. Data yang 100% unik

Data yang 100% unik, artinya dalam kolom tersebut tidak ada data yang sama. Kolom seperti ini tidak memberikan pengaruh terhadap hasil akhir. Biasanya berupa nomer untuk pengganti identitas. Contoh-contoh data unik 100%: Nomer Urut Pasien, Nomer Induk Pegawai, Nomer Pokok Mahasiswa, dan yang sejenisnya.

Penanganan data unik seperti ini adalah dengan menghapus keseluruhan kolom dari dataframe.

2. Missing-value dan interpolasi median.

Mengatasi missing-value pada satu (atau lebih) kolom data bukan langkah yang sederhana. Beberapa hal harus dipertimbangkan baik-baik. Pertimbangan pertama adalah: berapa prosen missing-value dalam kolom terperiksa. Untuk menjawabnya, ada pengetahuan umum yang berlaku yaitu: jika prosentase missing-value lebih besar dari 40 persen, maka langkah terbaik adalah menghapus kolom terperiksa dari dataframe. Lalu, jika missing-value kurang dari 40 persen, maka missing-value diisi dengan teknik tertentu.

Penelitian ini menggunakan teknik interpolasi median untuk mengisi missing-value. teknik interpolasi median bekerja dengan cara mengambil nilai tengah dari keseluruhan kolom dan digunakan untuk mengisi missing-value

3. Anomali pada data.

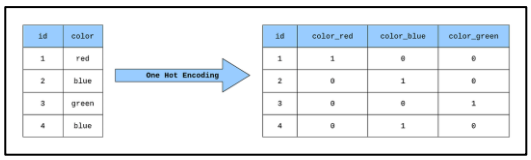
Banyak teknik yang dapat digunakan untuk mengatasi anomali dalam data. Khusus pada dataset yang digunakan pada penelitian ini, anomali ditemukan pada kolom "gender". Mayoritas dari kolom ini berisi "female" dan data male menjadi minoritas. Anomali ditemukan pada hanya satu data yang berisi "other". Langkah perbaikan yang terbaik adalah mengubah data "other" ke minoritas.

4. One-Hot Encoding (OHE).

Dalam ML, categorical-data harus diubah menjadi format numerik atau numerical-data, sebelum dapat digunakan untuk membangun model. Proses ini dikenal sebagai encoding. Beberapa metode yang bisa digunakan untuk encoding categorical-data, masing-masing memiliki kelebihan dan kekurangan tergantung pada konteks dan jenis data yang dihadapi.

One-Hot Encoding (OHE) adalah sebuah teknik konversi categorical-data menjadi format numerical-data sehingga dapat digunakan oleh algoritma ML. OHE mengubah categorical-data menjadi vektor biner. Setiap vektor memiliki panjang yang sama dengan jumlah kategori dan hanya memiliki satu nilai "1" di posisi yang sesuai dengan kategori tersebut, sementara semua posisi lainnya bernilai "0".

Perhatikan contoh berikut, diberikan sebuah categorical-data berisi data-data: "Red", "Green", dan "Blue". OHE akan mengubah data-data tersebut menjadi, perhatikan Gambar 2:



Gambar 2. Pengambaran OHE

- "Red" menjadi [1, 0, 0]
- "Green" menjadi [0, 1, 0]
- "Blue" menjadi [0, 0, 1]

5. SMOTEENN.

EDA memberikan hasil bahwa atribut keputusan (c.q. atribut stroke) dari dataset sangat tidak imbang (imbalanced dataset). Imbalanced data adalah suatu kondisi jumlah data class dalam dataset tidak memiliki distribusi yang sama. Misalkan dalam sebuah binary-classification, class A berjumlah lebih dari dua kali jumlah class B. Imbalanced Dataset, jika tidak diperbaiki, akan membuat algoritma ML membangun model yang cenderung bias pada majority class serta menurunnya akurasi dan performa model.

Penelitian menggunakan algoritm hybrid SMOTE+Edited Nearest Neighbors

(SMOTEENN), algoritma yang menggabungkan teknik Syntetic Minority Oversampling Technique (SMOTE) untuk membuat data sintetis pada minority class dan Edited Nearest Neighbors untuk membersihkan noisy data, class tambahan hasil dari SMOTE yang mungkin menjadi tumpang tindih dengan anggota majority-class.

6. Feature Scaling: Standard Scaller.

Feature scaling merupakan tahap terakhir dalam Data Preprocesing. Pada tahap ini, dilakukan normalisasi atau standarisasi terhadap data. Tujuannya adalah agar suatu independent variable menjadi sama pentingnya dengan independent variable yang lain.

Caranya dengan memproses kolom demi kolom dan disesuaikan menggunakan pola distribusi yang sama, sampai tidak ada independent variable yang bisa diabaikan.

7. Logistic Regression.

Logistic regression adalah metode analisis statistik yang digunakan untuk memodelkan hubungan antara variabel dependen biner (dikotomi) dan satu atau lebih variabel independen. Berbeda dengan regresi linear yang memprediksi nilai kontinu, logistic regression memprediksi probabilitas kejadian suatu peristiwa dengan menghasilkan output dalam bentuk nilai antara 0 dan 1. Fungsi logistik, atau sigmoid, digunakan untuk mengubah hasil prediksi regresi linear menjadi probabilitas, yang kemudian dapat diubah menjadi keputusan klasifikasi (misalnya, kelas 0 atau 1) berdasarkan ambang batas tertentu.

8. Confusion Matrix, Precision, Recall dan F1_Score.

a. Confusion Matrix

Confusion matrix adalah tabel yang digunakan untuk mengevaluasi kinerja model klasifikasi. Tabel ini menampilkan jumlah prediksi benar dan salah lalu dibandingkan dengan kelas aktual. Struktur dasarnya adalah:

True Positive (TP): Model memprediksi positif dan benar.

False Positive (FP): Model memprediksi positif tetapi salah.

True Negative (TN): Model memprediksi negatif dan benar.

False Negative (FN): Model memprediksi negatif tetapi salah.

b. Precision

Precision adalah rasio prediksi positif yang benar dibandingkan dengan total prediksi positif. Rumusnya adalah:

Precision = TP / (TP + FP)

Precision digunakan untuk mengukur seberapa tepat prediksi positif model, yaitu berapa banyak dari prediksi positif yang benar-benar positif.

c. Recall atau Sensitivity

Recall adalah rasio prediksi positif yang benar dibandingkan dengan total aktual positif. Rumusnya adalah:

Recall = TP / (TP + FN)

Recall berguna untuk mengukur seberapa baik model menemukan semua contoh positif yang sebenarnya.

d. F1_Score

F1_Score adalah nilai rata-rata harmonis dari Precision dan Recall. Rumusnya adalah:

F1_Score = 2 * ((Precision * Recall) / (Precision + Recall))

F1_Score sangat berguna terutama jika digunakan pada imbalanced dataset

HASIL DAN PEMBAHASAN

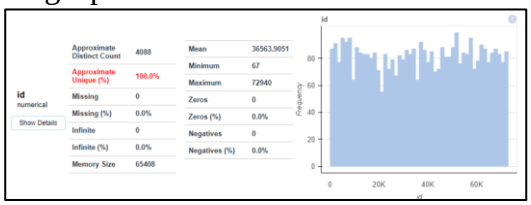
Hasil EDA dan Data Preprocessing.

Pada subbab ini ditampilkan hasil EDA beserta proses PreProcessing yang dilakukan terhadap kolom tersebut.

a. Kolom dengan useless-data.

Sebuah data dengan tingkat unik 100%, pada Gambar 3, berarti tidak ada lagi data yang sama pada keseluruhan

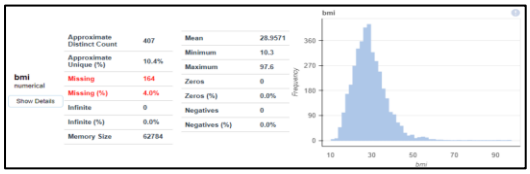
kolom tersebut. Pilihan terbaik adalah menghapus kolom id dari dataframe.



Gambar 3. Hasil EDA, kolom dengan Tingkat unik 100%

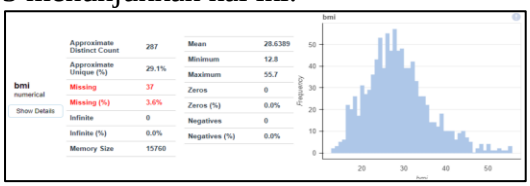
b. Kolom dengan missing-value.

Pada tahap EDA, ditemukan kolom bmi (body mass index), memiliki *missing-value* (Gambar 4). Jumlahnya, pada Training-data, ada sebanyak 164, atau empat prosen dari keseluruhan Training-data.



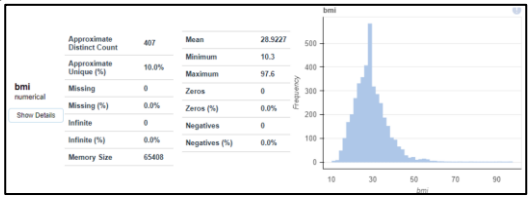
Gambar 4. Missing-value pada Training-Data

Missing-value pada test-data, sebanyak 37 data, atau tiga koma enam prosen dari keseluruhan test-data. Gambar 5 menunjukkan hal ini.



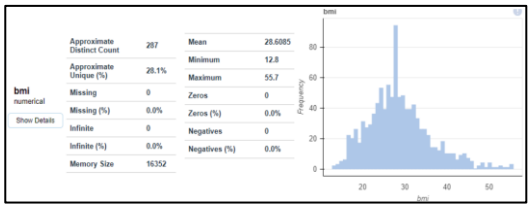
Gambar 5. Missing-value pada test-data.

Perbaikan yang dilakukan adalah dengan mengisi missing-value, pada Training-data dan Test-data, menggunakan teknik interpolasi median. Hasil perbaikan pada Training-data, dapat dilihat pada gambar 6.



Gambar 6. Hasil Perbaikan Missing-Value Pada Training-Data

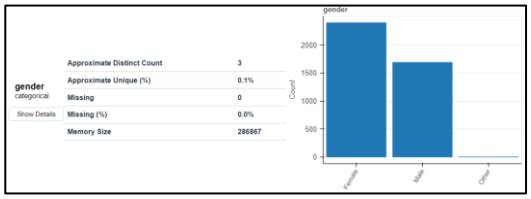
Perbaikan pada Test-data juga menggunakan teknik interpolasi median dengan pertimbangan, untuk menjaga konsistensi dengan Training-data. Hasil perbaikan dapat dilihat pada Gambar 7.



Gambar 7. Hasil perbaikan missing-valuepada Test-data

c. Kolom dengan outlier.

Anomali yang ditemukan pada proses EDA, hanya pada kolom gender dalam Training-data. Hasil observasi, Gambar 8, menunjukkan bahwa pada Training-data, kolom gender, mayoritas berisi “Female”. Kelas minoritas, terbagi menjadi dua, “Male” dan satu item data berisi “Others”.

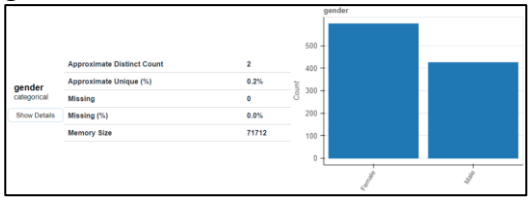


Gambar 8. Anomali pada kolom “gender”

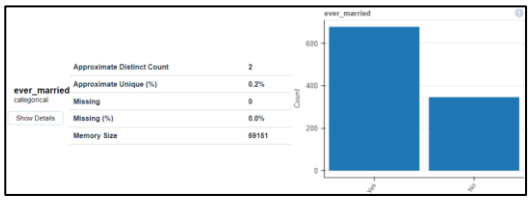
Perbaikan yang dilakukan adalah mengubah item “Others” menjadi “Male”.

d. Kolom dengan Categorical-data.

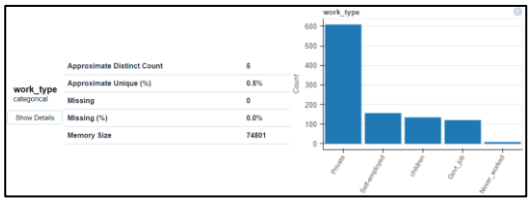
Algoritma Machine Learning membutuhkan data numerik untuk bekerja optimal. Proses EDA menunjukkan ada lima kolom dengan Categorical-data: “smoking_status”, “Residence_type”, “work_type”, “ever_married”, dan “gender”.



Gambar 9. Categorical-data pada kolom Gender



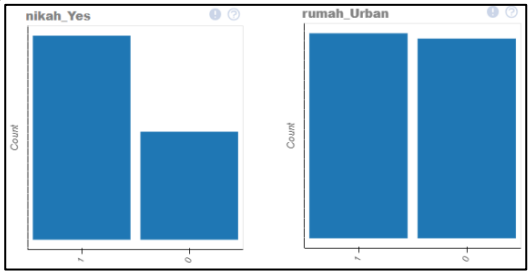
Gambar 10. Categorical-data pada kolom ever_married



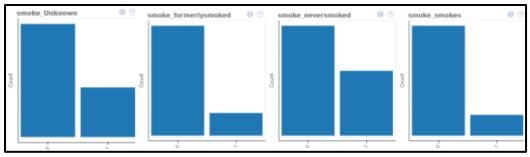
Gambar 11. Categorical-data pada kolom work_type

Gambar 12. Categorical-data pada kolom Residence type

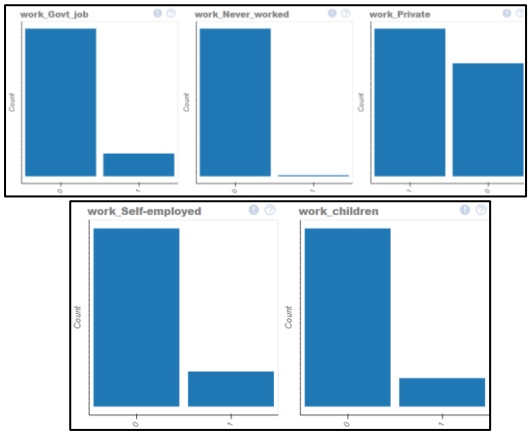
Teknik yang dipilih untuk konversi categorical-data menjadi numerical-data adalah One-Hot Encoding (OHE). Hasilnya dapat dilihat pada gambar-gambar berikut.



Gambar 13. Hasil OHE Pada Kolom Ever_Married Dan Residence Type



Gambar 14. Hasil OHE pada kolom ever_smoke

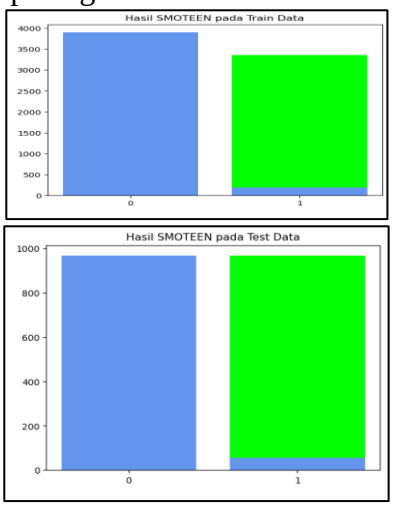


Gambar 15. Hasil OHE pada kolom work_type

e. Imbalanced Dataset.

Dataset yang imbalance, perbedaan jumlah antar class demikian besar sehingga ada majority-class dan minority-class. Perbedaan seperti ini, jika tidak diperbaiki, akan membuat model yang dibuat nantinya menjadi *overconfidence* dan cenderung bias pada majority-class.

Pada penelitian ini, digunakan teknik SMOTEENN, yang melakukan downsampling setelah proses oversampling selesai. Hasilnya dapat dilihat pada gambar 16.

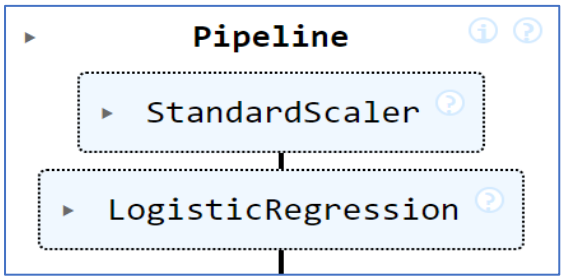


Gambar 16. Hasil SMOTEENN pada imbalanced train-data dan test-data

Gambar 16 dijelaskan sebagai berikut, sebelum proses SMOTEEN, training-data memiliki data class "0" sebanyak 3895 dan class "1" sebanyak 193 saja. Test-data memiliki class "0" sebanyak 966 dan class "1" hanya sebanyak 56 data. Setelah dilakukan proses SMOTEENN, traning-data sekarang memiliki class "1" sebanyak 3358 dan class "0" sebanyak 2808, sedangkan pada test-data sekarang, class "1" dan class "0" sama-sama memiliki 966 data, relatif seimbang.

Standarisasi dan klasifikasi.

Proses berikutnya adalah melakukan standarisasi pada train-data dan test-data. Meskipun data sudah di-split sebelum melakukan preprosesing, namun untuk memastikan bahwa tidak ada data-leaking, teknik pipelining digunakan dalam standarisasi dan pembangunan model.



Gambar 17. Pipelining Proses standarisasi dan klasifikasi

Hasil Percobaan.

Pada percobaan ini didapatkan hasil sebagai berikut:

Precision: 0.883.

Recall: 0.921.

F1 Score: 0.902

Precision: 0.883, hal ini berarti, dari semua prediksi yang dinyatakan positif oleh model, sebanyak 88.3 persen pasien memang benar-benar mengalami stroke. Nilai tersebut juga berarti, model menyatakan false-positive pada sebanyak 11.7 persen pasien.

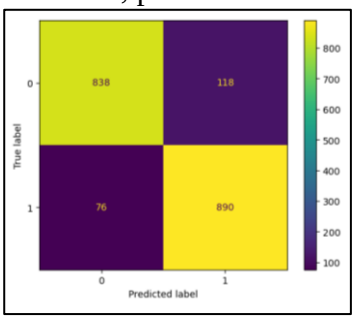
Recall: 0.921, berarti model berhasil mengidentifikasi 92.1 persen dari semua pasien yang benar-benar mengalami stroke.

Kekurangannya adalah, model gagal mengenali sebanyak 7.9 persen pasien yang benar-benar mengalami stroke. Nilai Recall yang tinggi memang diharapkan pada pembangunan model health_care seperti ini.

F1_score: 0.902, berarti model memiliki kombinasi yang baik antara antara kemampuan untuk mendeteksi semua true-positive dalam recall dan minimasi prediksi false-positive dari precision.

Nilai F1 Score menunjukkan bahwa model sangat efektif dan cukup akurat dalam prediksi.

Analisis pada model menggunakan confusion-matrix, pada Gambar 18.



Gambar 18. Confusion Matrix

Confusion-matrix pada Gambar 18 dijelaskan sebagai berikut:

True-Positive, sebanyak 890 pasien stroke, dideteksi menderita stroke. False-Positive, sebanyak 118 pasien sehat, dideteksi menderita stroke. True-Negative, sebanyak 838 orang sehat, dideteksi sehat. False-Negative, sebanyak 76 orang stroke, dideteksi sehat

SIMPULAN

Dataset didapat secara legal dari public repository kaggle. Penelitian berikutnya, mungkin perlu dipertimbangkan untuk mendapatkan dataset yang berisi data lokal Indonesia.

Dataset berhasil di-split secara fisik menjadi dua buah file .csv, dengan perbandingan 80:20. Pada penelitian berikutnya, sepertinya perlu diberikan variasi pada perbandingan splitting, misalkan 75:25 atau 90:10.

Proses EDA, sudah selesai dilakukan, dan memberikan hasil sejumlah data perlu dilakukan peningkatan atau perbaikan menggunakan preprocessing. Semua tahapan preprocessing yang diperlukan, baik pada train-data dan test-data, sudah selesai dilakukan.

Proses balancing-data menggunakan SMOTEEN, berhasil memperbaiki data yang tidak seimbang dengan baik. Perlu dipertimbangkan proses balancing-data lain sebagai bahan perbandingan.

Model sudah berhasil dibuat, dengan hasil precision sebesar 88,3 persen, hasil recall sebesar 92.1 persen dan F1_score sebesar 90.2 persen. Diperlukan penelitian lanjutan untuk melakukan analisis terhadap tingkat akurasi, dan penggunaan model ML lain sebagai pembanding.

DAFTAR PUSTAKA

Jinglin Peng, et.al, "DataPrep.EDA: Task-Centric Exploratory Data Analysis for Statistical Modeling in Python", https://arxiv.org/abs/2104.00841, SIGMOD 2021, Tanggal akses: 31/05/2024.

_____, The EU general data protection regulation (GDPR), “https://www.consilium.europa.eu/en/policies/data-protection/data-protection-regulation/”, accesed: 05/16/2024.

_____, California Consumer Privacy Act (CCPA), “https://oag.ca.gov/privacy/ccpa”, accesed: 05/16/2024.

Soriano, Federico, "Stroke Prediction Dataset", a public repository, https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset, tanggal akses: 31/05/2024