Pembelajaran Mesin: Penerapan Beberapa Metode Klasifikasi dan Metode Oversampling pada Dataset Stroke

Kevin Wijaya[1*], Hendy Ardhana[2], Allan Mateus[3]

[1*]Program Studi Teknik Informatika, Universitas Surabaya, Surabaya, Jawa Timur
[2]Program Studi Teknik Informatika, Universitas Surabaya, Surabaya, Jawa Timur
[3]Program Studi Teknik Informatika, Universitas Surabaya, Surabaya, Jawa Timur

Email: [1*]s160420136@student.ubaya.ac.id, [2]s160420138@student.ubaya.ac.id, [3]s160420071@student.ubaya.ac.id

(Naskah masuk: 14 Desember 2022, direvisi: dd mmm yyyy, diterima: dd mmm yyyy)

Abstrak

Pembelajaran mesin merupakan bidang ilmu yang berada di payung kecerdasan buatan. Bidang ini mempelajari mengenai algoritma komputer yang mengenali setiap pola yang ada pada data, dengan begitu komputer dapat semakin cerdas dengan belajar dari pengalaman data yang dimiliki. Berdasarkan teknik pembelajarannya, pembelajaran mesin dapat dibedakan menjadi *supervised learning dan unsupervised learning*. Pada penelitian atau eksperimen kali ini kami menggunakan teknik pembelajaran supervised learning yaitu klasifikasi. pada penelitian atau eksperimen yang kami lakukan pada dataset stroke, kami menyertakan beberapa metode lain seperti pra-pemrosesan data, pemilihan fitur, reduksi dimensi, model evaluasi, dan pengambilan sampel ulang. Terdapat empat jenis metode klasifikasi yang kami gunakan yaitu *k-Nearest Neighbors(KNN), Naive Bayes, Logistic Regression,* dan *Artificial Neural Network(ANN)*. dari setiap metode klasifikasi masing-masing memiliki tiga skenario yang berbeda. Tujuan menggunakan beberapa metode klasifikasi yang berbeda-beda serta menerapkan skenario yang berbeda-beda adalah untuk melihat perbandingan akurasi setiap metode klasifikasi dan skenario yang terjadi guna mengetahui metode klasifikasi dan skenario mana yang paling baik.

Kata Kunci: Pembelajaran Mesin, Klasifikasi, Oversampling, Stroke.

Machine Learning: The Application of Several Classification Methods and Oversampling Methods to the Stroke Dataset

Abstract

Machine learning is a field under the umbrella of artificial intelligence. This field studies computer algorithms that recognize every pattern in data, so that computers can be smarter by learning from the experience of the data they have. Based on learning techniques, machine learning can be divided into supervised learning and unsupervised learning. In this research or experiment, we used a supervised learning technique, namely classification. In the research or experiments we carry out on the stroke dataset, we include several other methods such as data pre-processing, feature selection, dimensionality reduction, evaluation models, and re-sampling. There are four types of classification methods that we use, namely k-Nearest Neighbors (KNN), Naive Bayes, Logistic Regression, and Artificial Neural Networks (ANN). Of each classification method each has three different scenarios. The purpose of using several different classification methods and applying different scenarios is to see a comparison of the accuracy of each classification method and the scenarios that occur in order to find out which classification method and scenario is the best.

Keywords: *Machine Learning, Classification, Oversampling, Stroke.*

I. PENDAHULUAN

A. Latar Belakang

Menurut WHO Stroke merupakan penyakit kedua mematikan di dunia. Stroke merupakan penyakit gangguan pembuluh darah di otak yang dipengaruhi oleh banyak hal. Salah satunya adalah penyumbatan pada sirkulasi pembuluh darah sehingga dapat menimbulkan pecahnya pembuluh darah di otak dan oksigen juga tidak dapat mencapai jaringan otak. Tanpa oksigen di jaringan otak, maka dapat menimbulkan sel-sel di otak mati dan mengakibatkan cacat fisik dan mental. Maka dari itu pentingnya menjaga aliran darah dan oksigen di otak. Untuk mencegah dan segera mengobati hal tersebut sejak dini maka jurnal ini dibuat untuk memprediksi apakah seseorang tersebut terkena stroke atau tidak.

B. Landasan Teori

1. Dataset: Stroke
   Stroke Dataset merupakan dataset yang terdiri dari 5110 observasi riwayat medis patient dan 12 atribut. atribut tersebut terdiri dari *id, gender, age, hypertension, heart_disease, ever_married, work_type, residence_type, avg_glucose_level, bmi, smoking_status*, dan *stroke*. class yang dimiliki oleh data target adalah class no dan class yes ditulis dalam bentuk binary 0 dan 1.

2. Pre-processing
   Pre-processing merupakan salah satu step yang ada di machine learning. Tujuannya adalah untuk memastikan apakah data memiliki kualitas yang baik agar ketika masuk proses modeling hasil yang diberikan adalah sangat baik. Terdapat beberapa jenis metode preprocessing yang digunakan antara lain yaitu handling missing value with median, encoding categorical, drop outliers, standardization scaling, min max scaling, dan normalization.

3. Feature Selection
   Feature Selection merupakan salah satu step yang ada di machine learning. Tujuannya adalah untuk menghapus features yang berulang dan tidak relevan dari dataset. sehingga waktu yang dibutuhkan untuk mengeksekusi model classification dapat lebih cepat dan efisien serta dapat meningkatkan akurasi.

4. Dimensionality Reduction
   Dimensionality Reduction merupakan salah satu bagian dari machine learning. Tujuannya adalah untuk mereduksi fitur atau atribut menjadi beberapa dimensi baru yang lebih kecil ukurannya tanpa menghilangkan arti dari fitur yang lama. Sama seperti feature selection, keuntungan menggunakan dimensionality reduction ini adalah waktu eksekusi yang diperlukan akan lebih cepat dan efisien serta dapat meningkatkan akurasi.

5. Oversampling
   oversampling merupakan salah satu metode resampling yang ada di statistik. Tujuan dari oversampling adalah untuk memperbaiki dataset yang data response atau target tidak seimbang(Imbalance). class target yang memiliki jumlah paling banyak disebut mayoritas dan class target yang memiliki jumlah paling sedikit disebut minoritas dengan menggunakan metode oversampling maka data minoritas akan ditingkatkan sebanyak data mayoritas sehingga setelah resampling data target akan seimbang.

6. Classification : k-Nearest Neighbors
   k-Nearest Neighbors (KNN) merupakan salah satu algoritma classification yang paling sederhana dengan membandingkan banyaknya tetangga yang paling dekat. KNN memprediksi class observasi baru dengan melihat persentase terbesar dari class observasi terdekat sebesar *k.*

7. Classification : Naive Bayes
   Naive Bayes merupakan salah satu algoritma classification yang menggunakan probabilitas classifier berdasarkan teorema bayes. Naive Bayes memprediksi observasi baru menggunakan probabilitas. Probabilitas digunakan untuk menghitung apakah suatu observasi baru masuk ke dalam suatu class. dimana class yang memiliki probabilitas paling tinggi yang akan dipilih.

8. Classification : Logistic Regression
   Logistic Regression merupakan salah satu algoritma classification yang menggunakan fungsi sigmoid untuk menghitung probabilitas dari suatu class observasi. Logistic Regression cocok untuk memprediksi target berbentuk binary classification.

9. Classification : Artificial Neural Network
   Artificial Neural Network (ANN) merupakan algoritma machine learning yang menjadi dasar dari algoritma deep learning. classifier ini mampu untuk menyelesaikan permasalah regression dan classification di beberapa bidang. Dari model perhitungannya, ANN hampir sama dengan logistic Regression namun di modifikasi. ANN mengandalkan neuron di beberapa layer untuk menghitung setiap nilai dari parameter berdasarkan fungsi aktivasi yang telah ditetapkan di layer tersebut.

10. Model Evaluation
    Model Evaluation merupakan salah satu step di machine learning. Model evaluation adalah untuk mengukur seberapa baik perfomance yang diberikan oleh suatu model dari classification itu. Tujuannya

untuk mengevaluasi performa model pada data-data yang belum terlihat sebelumnya.

C. Rumusan Masalah

- Apa jenis pre-processing yang digunakan?
- Apa jenis feature selection yang digunakan?
- Apa jenis dimensionality reduction yang digunakan?
- Apa jenis evaluation yang digunakan?
- Mengapa menggunakan metode oversampling?
- Mengapa menggunakan jenis classification tersebut?
- Skenario apa saja yang digunakan?
- Apakah skenario tersebut menghasilkan akurasi yang baik?

D. Tujuan

Tujuan dari penelitian ini adalah untuk membuat model machine learning yang dapat digunakan untuk memprediksi apakah suatu patient atau seseorang yang mempunyai riwayat medis tertentu sedang terkena penyakit stroke atau tidak. lalu untuk membuat model machine learning yang baik, maka kita melakukan beberapa rangkaian percobaan dan skenario untuk melihat model klasifikasi mana yang paling baik.

II. BAHAN DAN METODE

A. Bahan

Pada penelitian ini menggunakan bahan atau bisa di sebut sebagai dataset dari sebuah Halaman Website https://www.kaggle.com/ yaitu Stroke Prediction Dataset.

1. link source : https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
2. link drive (view) : https://drive.google.com/file/d/1nUnX4BJpVoFCFy4JXqUmCRqyuUWZTWuv/view?usp=share_link
3. Direct Link (downloadable) : https://drive.google.com/uc?export=download&id=1nUnX4BJpVoFCFy4JXqUmCRqyuUWZTWuv

B. Metode

Metode yang digunakan pada penelitian ini adalah :

1. Classification : k-Nearest Neighbors
2. Classification : Naive Bayes
3. Classification : Logistic Regression
4. Classification : Artificial Neural Network

C. Alur Penelitian

berikut merupakan alur dari penelitian atau eksperimen yang kami lakukan pada dataset stroke.
https://drive.google.com/file/d/1XisC6NriRINCiF5mPPYVWsKGZ28p9TY7/view?usp=sharing

Project UAS Stroke.diagram.png

III. HASIL DAN DISKUSI

Berikut merupakan tabel rangkuman dari setiap model dan model evaluasi terbaik berdasarkan original data dan resampling data.

Metode	Accuracy	
	Original Data	Resampling Data
KNN	95.1076%	95.6119%
Naive Bayes	93.8682%	77.3055%
Logistic Regression	95.1729%	77.2026%
ANN	94.1944%	91.6695%

Tabel 1. Hasil Modelling

Metode	Classification Report Accuracy	
	Original Data	Resampling Data
KNN	95%	96%
Naive Bayes	94%	77%
Logistic Regression	95%	77%
ANN	94%	92%

Tabel 2. Hasil Model Evaluation

Metode	K-Fold (k=30) Accuracy	
	Original Data	Resampling Data
KNN	95.1279%	92.6865%
Naive Bayes	86.0477%	76.8366%
Logistic Regression	95.1082%	76.435%
ANN	95.1082%	82.2472%

Tabel 3. Hasil Model Evaluation

A. Classification : k-Nearest Neighbors:

1. Original Data

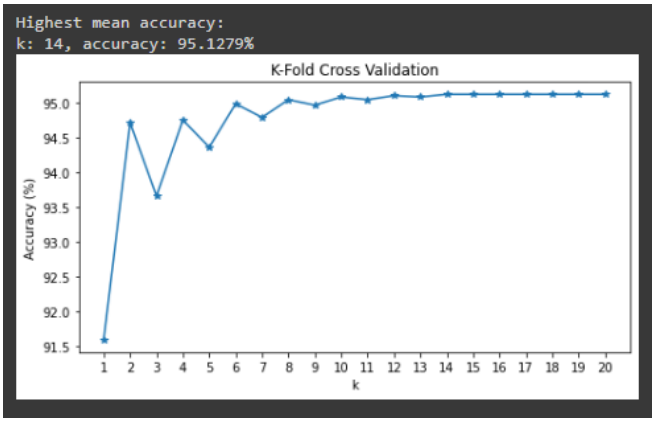
Pada original data akurasi terbaik diperoleh dengan persentase sebesar 95.1076%. Dengan proses di scaling menggunakan standarisasi, feature selection menggunakan anova dengan P-value < 10^{-5}, direduski menggunakan PCA dengan PC=1, kemudian k yang dipilih adalah 6.

```
Accuracy: 95.1076%
```

Gambar 1. Hasil accuracy KNN (Original Data)

```
Confusion Matrix:
[[1458    0]
 [  75    0]]

Normalized Confusion Matrix:
[[1. 0.]
 [1. 0.]]

Classification Report :
              precision    recall  f1-score   support

          no       0.95      1.00      0.97      1458
         yes       0.00      0.00      0.00        75

    accuracy                           0.95      1533
   macro avg       0.48      0.50      0.49      1533
weighted avg       0.90      0.95      0.93      1533
```

Gambar 2. Hasil Confusion Matrix and Classification Report KNN (Original Data)



Gambar 3. Hasil K-Fold Cross Validation KNN dengan k=30 (Original Data)

2. Resampling Data

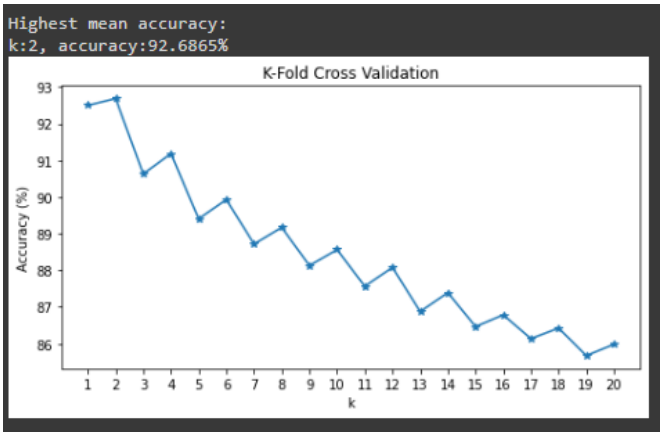
Pada resampling data akurasi terbaik diperoleh dengan persentase sebesar 95.6119%. Dengan proses di scaling menggunakan standarisasi, direduski menggunakan PCA dengan PC=9, kemudian k yang dipilih adalah 1.

```
Accuracy: 95.6119%
```

Gambar 4. Hasil accuracy KNN (Resampling Data)

```
Confusion Matrix:
[[1345  113]
 [  15 1444]]

Normalized Confusion Matrix:
[[0.92249657 0.07750343]
 [0.01028101 0.98971899]]

Classification Report :
              precision    recall  f1-score   support

          no       0.99      0.92      0.95      1458
         yes       0.93      0.99      0.96      1459

    accuracy                           0.96      2917
   macro avg       0.96      0.96      0.96      2917
weighted avg       0.96      0.96      0.96      2917
```

Gambar 5. Hasil Confusion Matrix and Classification Report KNN (Resampling Data)



Gambar 6. Hasil K-Fold Cross Validation KNN dengan k=30 (Resampling Data)

B. Classification : Naive Bayes

1. Original Data

Pada original data akurasi terbaik diperoleh dengan persentase sebesar 93.8682%. Dengan proses data outliers di drop, kemudian di scaling menggunakan standarisasi.

```
Accuracy: 93.8682%
```

Gambar 6. Hasil accuracy Naive Bayes (Original Data)

```
Confusion Matrix:
[[1435   23]
 [  71    4]]

Normalized Confusion Matrix:
[[0.98422497 0.01577503]
 [0.94666667 0.05333333]]

Classification Report :
              precision    recall  f1-score   support

          no       0.95      0.98      0.97      1458
         yes       0.15      0.05      0.08        75

    accuracy                           0.94      1533
   macro avg       0.55      0.52      0.52      1533
weighted avg       0.91      0.94      0.92      1533
```

Gambar 7. Hasil Confusion Matrix and Classification Report Naive Bayes (Original Data)

```
Highest accuracy: 86.0477%
```

Gambar 8. Hasil K-Fold Cross Validation Naive Bayes dengan k=30 (Original Data)

2. Resampling Data
   Pada resampling data akurasi terbaik diperoleh dengan persentase sebesar 77.3055%. Dengan proses di scaling menggunakan standarisasi, direduski menggunakan PCA dengan PC=10.

```
Accuracy: 77.3055%
```

Gambar 8. Hasil Confusion Matrix and Classification Report Naive Bayes (Resampling Data)

```
Confusion Matrix:
[[ 975  483]
 [ 179 1280]]

Normalized Confusion Matrix:
[[0.66872428 0.33127572]
 [0.12268677 0.87731323]]

Classification Report :
              precision    recall  f1-score   support

          no       0.84      0.67      0.75      1458
         yes       0.73      0.88      0.79      1459

    accuracy                           0.77      2917
   macro avg       0.79      0.77      0.77      2917
weighted avg       0.79      0.77      0.77      2917
```

Gambar 9. Hasil Confusion Matrix and Classification Report Naive Bayes (Resampling Data)

```
Accuracy: 76.8366%
```

Gambar 10. Hasil K-Fold Cross Validation Naive Bayes dengan k=30 (Resampling Data)

C. Classification : Logistic Regression
   1. Original Data
      Pada original data akurasi terbaik diperoleh dengan persentase sebesar 95.1729%. Dengan proses di scaling menggunakan standarisasi. kemudian di feature selection menggunakan anova dengan P-value < 10^{-5}, kemudian di reduksi dengan PCA menggunakan PC=2.

```
Accuracy: 95.1729%
```

Gambar 11. Hasil accuracy Logistic Regression (Original Data)

```
Confusion Matrix:
[[1458    0]
 [  74    1]]

Normalized Confusion Matrix:
[[1.         0.        ]
 [0.98666667 0.01333333]]

Classification Report :
              precision    recall  f1-score   support

          no       0.95      1.00      0.98      1458
         yes       1.00      0.01      0.03        75

    accuracy                           0.95      1533
   macro avg       0.98      0.51      0.50      1533
weighted avg       0.95      0.95      0.93      1533
```

Gambar 12. Hasil Confusion Matrix and Classification Report Logistic Regression (Original Data)

```
Highest accuracy: 95.1082%
```

Gambar 13. Hasil K-Fold Cross Validation Logistic Regression dengan k=30 (Original Data)

2. Resampling Data
   Pada resampling data akurasi terbaik diperoleh dengan persentase sebesar 77.2026%. Dengan proses di scaling menggunakan standarisasi, kemudian dilakukan feature selection menggunakan anova dengan P-value < 10^{-50} direduski menggunakan PCA dengan PC=4

```
Accuracy: 77.2026%
```

Gambar 14. Hasil accuracy Logistic Regression (Resampling Data)

```
Confusion Matrix:
[[1048  410]
 [ 255 1204]]

Normalized Confusion Matrix:
[[0.71879287 0.28120713]
 [0.17477724 0.82522276]]

Classification Report :
              precision    recall  f1-score   support

          no       0.80      0.72      0.76      1458
         yes       0.75      0.83      0.78      1459

    accuracy                           0.77      2917
   macro avg       0.78      0.77      0.77      2917
weighted avg       0.78      0.77      0.77      2917
```

Gambar 15. Hasil Confusion Matrix and Classification Report Logistic Regression (Resampling Data)

```
Accuracy: 76.435%
```

Gambar 16. Hasil K-Fold Cross Validation Logistic Regression dengan k=30 (Resampling Data)

D. Classification : Artificial Neural Network

1. Original Data

   Pada original data akurasi terbaik diperoleh dengan persentase sebesar 94.1944%. Dengan proses data outliers di drop kemudian di scaling menggunakan standarisasi. dengan activation function di hidden layer adalah tanh, dengan hidden layer berjumlah 2 dan neuron di layer 1 berjumlah 9 dan neuron di layer 2 berjumlah 3.

```
activation: tanh, accuracy: 94.1944%
```

Gambar 17. Hasil accuracy ANN (Original Data)

```
Confusion Matrix:
[[1436   22]
 [  67    8]]

Normalized Confusion Matrix:
[[0.98491084 0.01508916]
 [0.89333333 0.10666667]]

Classification Report :
              precision    recall  f1-score   support

          no       0.96      0.98      0.97      1458
         yes       0.27      0.11      0.15        75

    accuracy                           0.94      1533
   macro avg       0.61      0.55      0.56      1533
weighted avg       0.92      0.94      0.93      1533
```

Gambar 18. Hasil Confusion Matrix and Classification Report ANN (Original Data)

```
Highest accuracy: 95.1082%
```

Gambar 19. Hasil K-Fold Cross Validation ANN dengan k=30 (Original Data)

2. Resampling Data

   Pada original data akurasi terbaik diperoleh dengan persentase sebesar 91.6695%. Dengan proses di scaling menggunakan standarisasi. dengan activation function di hidden layer adalah logistic, dengan hidden layer berjumlah 2 dan neuron di layer 1 berjumlah 300 dan neuron di layer 2 berjumlah 100.

```
Accuracy: 91.6695%
```

Gambar 19. Hasil accuracy ANN (Resampling Data)

```
Confusion Matrix:
[[1313  145]
 [  98 1361]]

Normalized Confusion Matrix:
[[0.9005487  0.0994513 ]
 [0.06716929 0.93283071]]

Classification Report :
              precision    recall  f1-score   support

          no       0.93      0.90      0.92      1458
         yes       0.90      0.93      0.92      1459

    accuracy                           0.92      2917
   macro avg       0.92      0.92      0.92      2917
weighted avg       0.92      0.92      0.92      2917
```

Gambar 20. Hasil Confusion Matrix and Classification Report ANN (Resampling Data)

```
Highest accuracy: 82.2472%
```

Gambar 21. Hasil K-Fold Cross Validation ANN dengan k=30 (Resampling Data)

IV. KESIMPULAN

Dari melakukan percobaan untuk mencari prediksi yang paling akurat dari keempat metode yang diambil dari resampling data dapat disimpulkan bahwa metode k-Nearest Neighbors (KNN) tertinggi dengan angka 95.6119%, lalu selanjutnya ada Artificial Neural Network (ANN) dengan angka 91.6695%, lalu Naive Bayes dengan angka 77.3055% sedangkan Logistic Regression paling kecil dengan angka 77.2026%. Waktu yang diperlukan untuk mengeksekusi seluruh code yang ada di google colab sebesar ± 20 menit. Sehingga disarankan untuk mengaktifkan mode GPU untuk mempercepat proses eksekusi.

REFERENSI

[1] Dios Kurniawan, M. S. (2022). Pengenalan Machine Learning dengan Python. Culemborg, Netherlands: Van Duuren Media.

[2] Lemaitre, G., Nogueira, F., & Aridas, C. K. (2017, 17 1). Imbalanced-learn: A Python Toolbox to Tackle the Curse of. Retrieved from Imbalanced-learn: A Python Toolbox to Tackle the Curse of: https://www.jmlr.org/papers/volume18/16-365/16-365.pdf

[3] Gambaran Faktor Risiko Dan Tipe stroke pada Pasien Rawat Inap Di Bagian Penyakit Dalam RSUD Kabupaten Solok Selatan Periode 1 Januari 2010 - 31 Juni 2012 | Dinata | Jurnal Kesehatan Andalas. (n.d.). https://jurnal.fk.unand.ac.id/index.php/jka/article/view/119