Jurnal Sains, Nalar, dan Aplikasi Teknologi Informasi Vol. 1 No. 1 (2021)

E-ISSN: 2807-5935 DOI: 10.20885/snati.v1i1.2

PENANGANAN IMBALANCE DATA PADA KLASIFIKASI KEMUNGKINAN PENYAKIT STROKE

Siti Mutmainah
Program Studi Informatika
Universitas Islam Indonesia
Jl. Kaliurang 14,5 Sleman,
Yogyakarta, Indonesia 55584
20917057@students.uii.ac.id

ABSTRAKSI

Penanganan *imbalance data* dilakukan untuk menangani distribusi data yang tidak imbangnya antar *class* mayoritas dan *class* minoritas, hal ini dapat mengakibatkan *Machine Learning* keliru dalam melakukan klasifikasi. Aritkel ini menyajikan reviu jurnal literatur yang pernah dilakukan dalam kurun tahun 2013-2021 dalam penanganan *imbalance data*. Penanganan i*mbalance data* digunakan agar distribusi antar data *class* mayoritas dan *class* minoritas menjadi seimbang. Metode, akurasi atau performa dan dataset yang digunakan oleh literatur dalam klasifikasi akan dibandingkan. Berdasarkan reviu, performa yang dihasilkan cukup beragam dan terdapat beberapa teknik yang dapat dilakukan untuk menangani *Imbalance data.* Teknik-teknik tersebut bekerja dengan mendorong *class* minoritas mengubah distribusi *class* minoritas dan mayoritas. Teknik yang digunakan yaitu *Random oversampling* dan *Random undersampling* pada kemungkinan penyakit stroke. Penanganan *Imbalance data* dilakukan antar *class* 1 (Stroke) dan *class* 0 (tidak stroke) dengan distribusi antar data sama. Hasil yang didapatkan pada penggunaan teknik *random oversampling* mendapat performa yang lebih tinggi yaitu 95% daripada teknik r*andom undersampling* yang mendapat performa 76%.

Kata Kunci

Imbalance Data; Classification; Data Balancing, Penyakit Stroke.

1. PENDAHULUAN

Serebrovaskular atau lebih dikenal dengan (Stroke) merupakan penyakit yang menyebabkan kematian No.2 dan penyebab utama No.3 dari kecacatan diseluruh dunia, Stroke terjadi akibat penyumbatan atau pecahnya arteri ke otak [1]. Otak adalah organ yang kompleks yang dapat mengontrol fungsi tubuh. Saat terkena Stroke aliran darah tidak mencapai otak sehingga tidak dapat berfungsi sebagaimana mestinya [2].

Faktor resiko stroke meliputi usia, jenis kelamin, hipertensi, penyakit jantung, obesitas, rokok serta kurangnya aktifitas fisik. Stroke dapat dicegah dengan menerapkan perilaku hidup sehat seperti berolahraga dan menghindari makanan yang berkolesterol tinggi dan tidak merokok [3].

Hasil Riset Kesehatan Dasar (Riskendas) penyakit tidak menular di Indonesia, jumlah kematian pada semua umur akibat stroke mencapai 15,4%. Kematian di kelompok usia 45-54 tahun yang tinggal pada daerah perkotaan 15,9%, sedangkan pada daerah pedesaan 11,5%. Kematian kelompok usia 55-64 tahun yang tinggal di perkotaan 26,8%, pada pedesaan 17,4%. Berdasarkan survey, kasus stroke di indonesia makin tinggi tiap tahunnya [4].

Dapat disimpulkan bahwa stroke merupakan penyakit yang mengancam kehidupan hingga menyebabkan kematian dan dapat menyebabkan kecacatan seperti disebutkan [1] [2] [3] [4]. Ini menjadi patokan untuk mewaspadai adanya kemungkinan terkena stroke, oleh karena itu perlu mempelajari adanya kemungkinan terjadinya stroke dari variabel-variabel yang didapat pada data pasien stroke. Data tersebut akan diklasifikasi menggunakan *Machine learning*. Klasifikasi adalah teknik mempelajari kumpulan data sehingga menghasilkan aturan yang dapat mengenali data baru yang belum dipelajari, Klasifikasi dapat didefinisikan sebagai proses pemetaan objek data menjadi kategori dari salah satu *class* yang telah ditentukan sebelumnya [5].

Distribusi *Imbalance data* antar *class* merupakan masalah yang cukup umum terjadi pada data medis. Data yang tidak imbang antar *class* mayoritas dan minoritas dapat menyebabkan kekeliruan dalam melakukan klasifikasi. Data yang mengalami *Imbalance data* antar *class* akan mengandalkan *class* mayoritas dalam pengklasifikasi. Konsekuesi dari diagnosis yang dilakukan adalah mendapatkan hasil klasifikasi yang salah dan menyebabkan pasien salah penanganan. Mempelajari masalah *Imbalance data* antar *class* sangat penting dalam sudut pandang dunia medis [6]. Artikel ini menyajikan reviu literatur pada penanganan *Imbalance data* dan mempelajari metode serta teknik yang digunakan dalam penelitian terdahulu. Selain itu reviu ini bertujuan untuk melihat performa teknik penanganan *Imbalance data* dan ragam klasifikasi yang pernah dilakukan. Hasil reviu akan digunakan untuk menentukan teknik penanganan *Imbalance data* yang dapat digunakan pada klasifikasi kemungkinan penyakit stroke.

2. SELEKSI LITERATUR

Strategi yang dilakukan untuk mencari referensi literatur yang digunakan berasal dari mesin pencarian: Google Scholar, Science Direct dan jurnal serta naskah publikasi skripsi maupun tesis pada Universitas Islam Indonesia (https://library.uii.ac.id/osr).

Seleksi literatur terbatas hanya pada penanganan atau penggunaan teknik *Data balancing* pada klasifikasi saja. Kata Kunci yang dipakai dalam pencarian literatur adalah: *Imbalance Data, Classification, Data Pre-processing, Data Balancing.* Kata kunci ditranslasikan ke bahasa Indonesia dan Bahasa Inggris. Literatur yang didapatkan berbahasa Indonesia dan Inggris. Literatur yang sesuai dengan pencarian kata kunci berjumlah 9 literatur dan yang digunakan adalah yang terbit antara tahun 2013-2021.

Tabel 1. Fokus setiap literatur

No	Referensi	Jumlah
1	Literatur 1	Segmentasi Customer pada Industri Perbankan
2	Literatur 2	Prediksi Customer churn
3	Literatur 3	Klasifikasi Objektivitas Berita Online
4	Literatur 4	Klasifikasi Penilaian Kredit
5	Literatur 5	Klasifikasi kelayakan donor darah
6	Literatur 6	Noise identification noise identification in imbalanced data
7	Literatur 7	KEBAKARAN HUTAN
8	Literatur 8	level analysis for cancer diagnosis
9	Literatur 9	Classifier ensemle desain for imbalance data

Tabel 2. Penggunaan metode

No	Metode	Jumlah
1	Naïve Bayes	4
2	K-Nearest Neighbors	4
3	C4.5	3
4	SVM	2
5	Random Forest	2
6	NN	1
7	CNN	1
8	MLP	1
9	RIPPER	1

Tabel 3. Teknik *Data Balancing*.

No	Solusi	Jumlah
1	SMOTE	6
2	SMOTE-Oversampling	2
3	SMOTE-Undersampling	1
4	Random Oversampling	1
5	SMOTE-LOF	1
6	SMOTE-ENN	1
7	SMOTE-TL	1
8	SPIDER	1
9	SPIDER2	1
10	Hybrid Approach	1

Tabel 4. Jumlah dataset

No	Referensi	Jumlah
1	Literatur 1	200/17
2	Literatur 2	12
3	Literatur 3	200/13
4	Literatur 4	232
5	Literatur 5	247/9
6	Literatur 6	768
7	Literatur 7	502/6
8	Literatur 8	-
9	Literatur 9	1916

3. STUDI LITERATUR

Studi literatur dilakukan untuk mempelajari serta melihat pola dan pengelompokan dari literatur reviu yang digunakan. pengelompokan dilakukan untuk melihat metode klasifikasi yang digunakan, solusi penanganan *Imbalance data* serta fokus klasifikasi yang dilakukan masing-masing literatur.

3.1 Metode dalam Literatur

Naïve Bayes dan K-Nearest Neighbors adalah metode klasifikasi yang paling banyak digunakan dalam literatur. Tabel 2 menunjukan beberapa metode dan jumlah penggunaannya pada literatur.

3.2 Solusi dalam Literatur

Solusi pada masing-masing literatur dalam menyelesaikan masalah *imbalance data* sangat beragam. Diantaranya adalah SMOTE (*Syntetic Minority Over Sampling Technique*) diantaranya [7][8][9], kemudian penggunaan K-means [10] dan Hybrid Approach [11]. SMOTE menangani *class* yang tidak seimbang pada dataset. *Class* yang tidak seimbang merupakan *class* dengan mayoritas lebih besar dari yang minoritas. Konsep SMOTE yaitu membuat replica dari *class* minoritas sehingga mirip dengan *class* mayoritas, *replica* ini disebut sintetis [7]. Komparasi solusi penanganan *Imbalance data* juga disajikan untuk perbandingan tiap teknik *Data Balancing* [6][12][13][14]. Beberapa teknik *Data Balancing* yang dipakai untuk menyelesaikan masalah dapat dilihat pada Tabel 3.

3.3 Dataset Literatur

Dataset yang digunakan memiliki jumlah record dan variabel yang berbeda. Klasifikasi yang dilakukan juga berbeda antara satu sama lain literatur. Jumlah record dan variabel dataset disajikan pada Tabel 4.

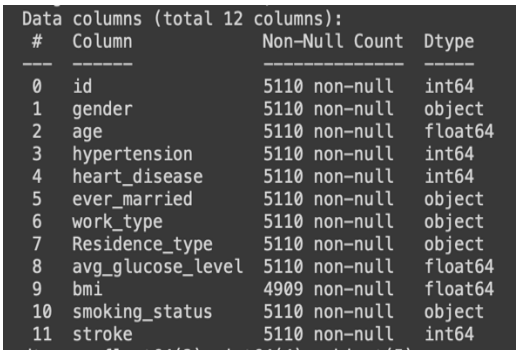
4. ANALISIS LITERATUR

Melalui reviu didapatkan beberapa teknik yang dapat dilakukan untuk penanganan *Imbalance data.* Teknik-teknik ini diterapkan untuk mendorong *class* minoritas serta mengubah distribusi *class* minoritas dan mayoritas. Performa yang dihasilkan pada beberapa literatur cukup beragam dikarenakan perbedaan proporsi dataset yang digunakan oleh masing-masing literatur. Performa yang didapat akan semakin rendah dengan penggunaan dataset yang cukup besar, oleh karena itu perlu dilakukan pemilihan metode sesuai dengan dataset yang digunakan. Pemilihan teknik *Data Balancing* akan disesuaikan dengan *Pre-processing* yang dilakukan. Tabel 5 menyajikan beberapa teknik yang dapat diterapkan [6].

Dari teknik-teknik *Data Balancing* yang digunakan serta pada literatur, teknik SMOTE merupakan yang paling banyak dipakai dalam penanganan *Imbalance* serta komparasi teknik yang dapat digunakan pada literatur reviu. Teknik Oversampling mendapat performa yang paling tinggi dibandingkan teknik Undersampling dan SMOTE [12]. Pada komparasi lain, teknik Random-Oversampling memiliki performa lebih tinggi yaitu 97% daripada teknik SMOTE-Oversampling yaitu 87.33% [13]. Kedua penelitian tersebut menggunakan metode yang berbeda namun dengan jumlah dataset yang tidak jauh berbeda yaitu 232 dan 247. Literatur lainya melakukan komparasi SMOTE dan SMOTE-LOF, SMOTE-LOF mendapat performa paling tinggi pada metode C4.5. Pada metode Naïve Bayes dan SVM performa tertinggi didapatkan oleh teknik SMOTE. Pada literatur juga ditemukan bahwa jumlah dataset mempengaruhi performa yang didapatkan dalam klasifikasi.

Tabel 5. Macam-macam Teknik *Balancing*

No	Referensi	Jumlah
1	SMOTE	Menghasilkan data sintetis, sesuai dengan jumlah data mayoritas.
2	Oversampling	Menandakan *class* minoritas melalui pembuatan sampel baru atau mengulangi yang lama.
3	Undersampling	Meratakan sampel dari kedua *class* dengan mengurangi data dan membuang sampel dari *class* mayoritas.
4	Random Oversampling	Mengambil secara acak dan menghasilkan sampel baru untuk *class* minoritas sehingga kedua *class* memiliki jumlah sampel yang sama.
5	SMOTE-TL	Menghapus sampel yang menyusun tautan Tomek.
6	SMOTE-ENN	Gabungan dari pengembangan SMOTE by ENN untuk memfilter derau.
7	SPIDER	Menggabungkan oversampling *class* minoritas dan memfilter sampel keras dari *class* mayoritas.
8	SPIDER2	Membagi menjadi dua tahap termasuk pra-pemrosesan sampel *class* mayoritas dan minoritas.
9	ADASYN	Menggunakan distribusi berbobot untuk sampel yang berbeda dari *class* minoritas.
10	ADOMS	Menghasilkan sampel sintetis di sepanjang prinsip pertama sumbu komponen data.
11	Borderline-SMOTE	Mencapai prediksi dengan mempelajarinya hingga akurasi yang didapatkan maksimum.
12	Safe-Level-SMOTE	Sebelum membuat sampel sintetis, teknik ini menentukan tingkat aman dari sampel *class* minoritas ac-cording ke SMOTE.
13	Metode hibrida	Menggabungkan dua teknik resampling.



```
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
```

Gambar 1. Variabel dataset.

Semakin banyak dataset yang dipakai semakin rendah performa klasifikasi. Oleh karena itu, teknik penanganan *Imbalance data* yang akan digunakan pada klasifikasi kemungkinan penyakit stroke adalah Random-Oversampling dan Random Undersampling. Teknik ini digunakan karena memiliki akurasi yang tinggi.

5. IMPLEMENTASI

Setelah melakukan reviu literatur penanganan *Imbalance Data* yang membahas penggunaan teknik *Data Balancing* pada klasifikasi, pemodelan dilakukan dengan menggunakan metode klasifikasi Random Forest. Penanganan *Imbalance Data* menggunakan teknik Random-Oversampling dan Random-Undersampling. Teknik akan dilakukan untuk mengecek akurasi sebelum dan sesudah penerapan *Data Balancing.* Setelah melakukan Training data selanjutnya menerapkan model dengan data Testing. Penelitian akan menggunakan Bahasa pemprograman Python untuk mengolah data dan menggukan jupyter sebagai notebook.

5.1 Dataset

Dataset yang digunakan adalah data dari kaggle.com. *Stroke Prediction Dataset* dengan 12 Variabel dan 5110 *Record.* Variabel terdiri dari: *Id, gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status, stroke*. Variabel yang ada akan dikorelasikan untuk melihat hubungan yang ada antar variabel melalui visualisasi data. Lihat Gambar 1.

- Variabel gender: Variabel ini terdiri dari 3 kategori yaitu Female lebih banyak yaitu 2994 orang, male berjumlah 2115 orang dan other berjumlah 1 orang.
- Variabel age: Variabel ini berisi umur pasien.
- Variabel heart_disease: Variabel ini berisi apakah pasien tersebut memiliki penyakit jantung atau tidak. kategori No lebih banyak yaitu 4834 orang, kategori Yes berjumlah 276 orang dan other berjumlah 1 orang.
- Variabel hypertension: Variabel ini berisi apakah pasien tersebut memilik penyakit hypertensi ketegori 0 (Tidak) 4612 dan kategori 1 (iya) 498Variabel ini berisi umur pasien.
- Variabel Residence_type: Variabel ini berisi 2 kategori Urban 2596 dan Rural 2514.
- Variabel ever_married: Variabel ini terdiri dari 2 kategori yaitu kategori No 1757orang dan kategori Yes berjumlah 3353 orang.
- Variabel work_type: Variabel ini berisi Private 2925, self-employed 819, children 687, Govt_job 657, Never_worked 22.
- Variabel smoking_status: Variabel ini berisi Never smoke 1892, Unknow 1544, formery smoked 885, smoked 789.
- Variabel avg_glucose_level: Variabel ini berisi kadar glukosa pada pasien.
- Variabel stroke: Variabel ini berisi 2 kategori 1 (stroke) 249 dan 0 (tidak) 4861.

Dataset yang ada memiliki *class imbalance data* antara 1 (Stroke) dan 0 (tidak stroke). Masalah seperti ini cukup umum pada data medis. Dataset yang mengalami *imbalance* antar *class* mayoritas dan minoritas dapat penyebabkan kekeliruan dalam melakukan klasifikasi karena mengandalkan *class* mayoritas. *Imbalance data* antar *class* dapat dilihat pada Gambar 2**.**

5.2 Deskripsi Variabel

Deskripsi dilakukan untuk melihat visualisasi jumlah data antar variabel serta mempelajari variabel yang ada sebelum melakukan praprosessing dan pemodelan. Variabel yang digunakan: *gender, age, hypertension, heart_disease, ever_married, work_type, residence_type, avg_glucose_level, bmi, smoking_status, stroke.*

5.2.1 Korelasi antar variabel

Korelasi antar variabel yang ada terlihat pada Gambar 3**.** Semakin gelap warna diagram semakin tinggi tingkat korelasi antar variabel yang ada.

5.2.2 Kasus stroke berdasarkan gender

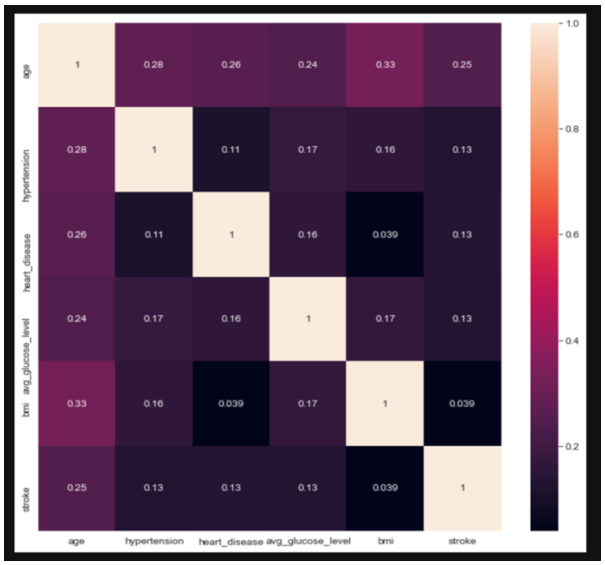
Visualisasi kasus stroke berdasarkan *gender* ditujukan untuk melihat jumlah dataset yang mengalami stroke berdasarkan gender. Pada dataset ini ditemukan bahwa pasien yang mengalami stroke dengan *gender female* lebih banyak dari pada *gender male,* lihat Gambar 4**.**

5.2.3 Distribusi stroke berdasarkan umur

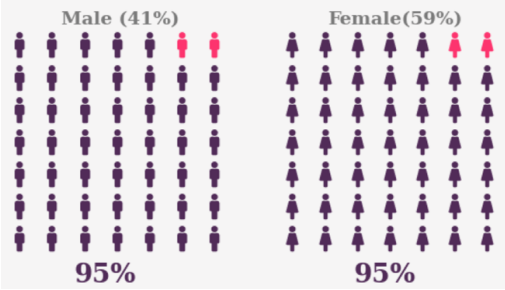
Visualisasi ini dilakukan untuk melihat rentang umur dari jumlah data pasien yang terkena penyakit stroke. Semakin tua umur maka semakin meningkat kemungkinan terkena stroke, hal ini terlihat pada diagram berdasarkan class 1 dan 0, lihat Gambar 5.

5.2.4 Distribusi stroke berdasarkan heart_disease

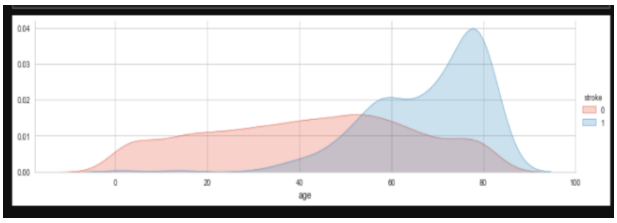
Visualisasi ini dilakukan untuk melihat jumlah kasus stroke berdasarkan variabel *heart_disease* atau penyakit jantung dari masing-masing *class* 1 dan 0. Distribusi data No atau tidak



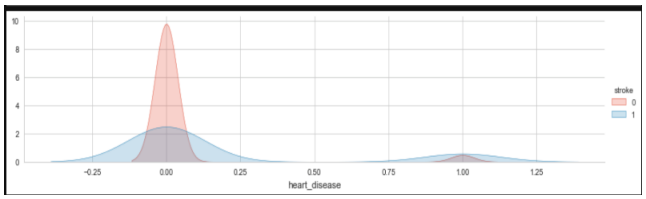
Gambar 3. Korelasi antar variabel.



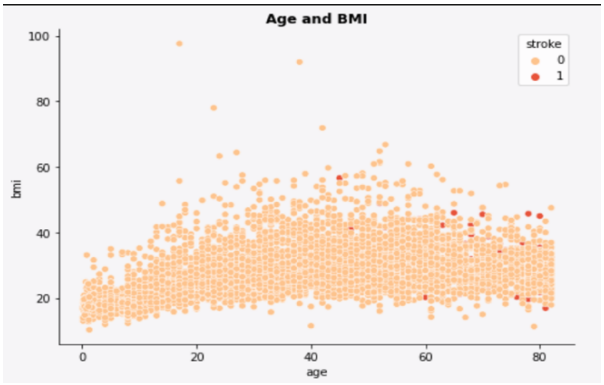
Gambar 4. Variabel gender stroke.



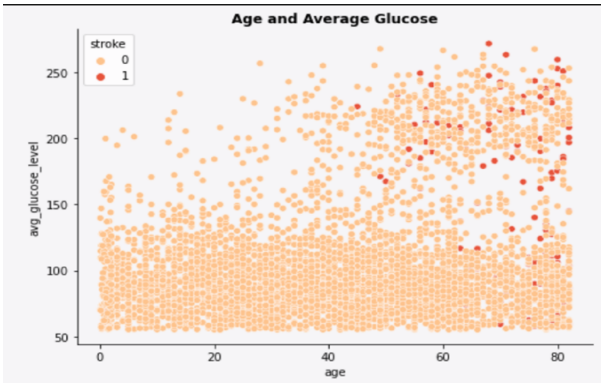
Gambar 5. stroke berdasarkan age.



Gambar 6. stroke berdasarkan heart_disease.



Gambar 7. Age dan BMI.



Gambar 8. Age dan *avg_glucose_level*

mengalami penyakit jantung lebih besar, terlihat pada bagian kiri diagram pada Gambar 6. Pasien yang tidak mengalami penyakit jantung lebih besar kemungkinan untuk tidak mengalami stroke.

5.2.5 Distribusi stroke berdasarkan age dan BMI

Visualisasi ini melihat jumlah kasus stroke berdasarkan variabel *age* dan BMI dari dataset. Gambar 7 menunjukkan sebaran data antar *class* 1 dan 0, sumbu X adalah umur pasien dan sumbu Y adalah BMI dari pasien pada dataset.

5.2.6 Distribusi stroke berdasarkan age dan glukosa

Visualisasi ini dilakukan untuk melihat jumlah kasus stroke berdasarkan variabel *age* dan *avg_glucose_level* pada dataset. Sumbu X adalah umur pasien dan sumbu Y adalah level glukosa

pasien pada dataset. Gambar 8 menunjukkan semakin tinggi level glukosa dan semakin tua umur pasien akan semakin tinggi kemungkinan untuk terkena stroke.

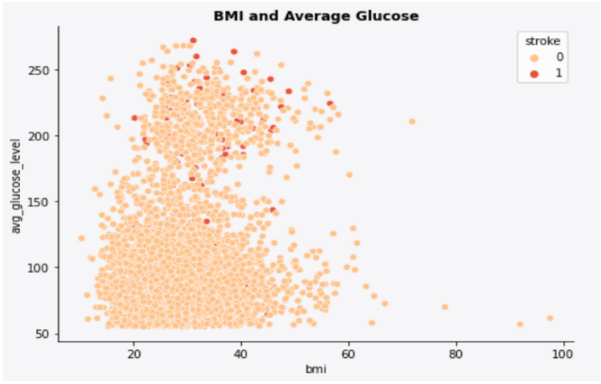
5.2.7 Stroke berdasarkan BMI dan glucose

Visualisasi ini dilakukan untuk melihat jumlah kasus stroke berdasarkan variabel BMI dan *avg_glucose_level.* Sumbu X adalah BMI pasien dan sumbu Y adalah level glukosa pasien pada dataset. Gambar 9 menunjukkan kadar glukosa yang tinggi akan meningkatkan potensi kemungkinan pasien mengalami stroke. Sedangkan pasien yang memiliki BMI antara 25 hingga 60 memiliki kecenderungan mengalami kemungkinan penyakit stroke.

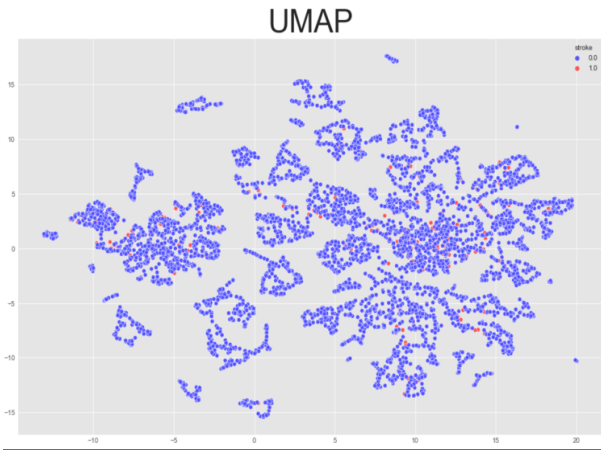
5.3 Exploratory Data Analysis

Setelah melihat visualisasi antar variabel serta deskripsi variabel yang ada. Tahap selanjutnya adalah *exploratory data analysis*. Tahap ini berguna untuk memahami data. Pada kasus ini dataset yang digunakan mempunyai 2 *class* dan dataset yang mengalami *imbalance* antar *class* mayoritas dan minoritas, hal ini menyebabkan *machine learning* akan keliru. Terbukti pada *class* 1 atau *class* minoritas tampak tidak dapat dibedakan dengan baik karena mengalami *imbalance data* antara *class* 1 dan 0. Hal ini dapat dilihat dari persebaran data yang ada pada Gambar 10.

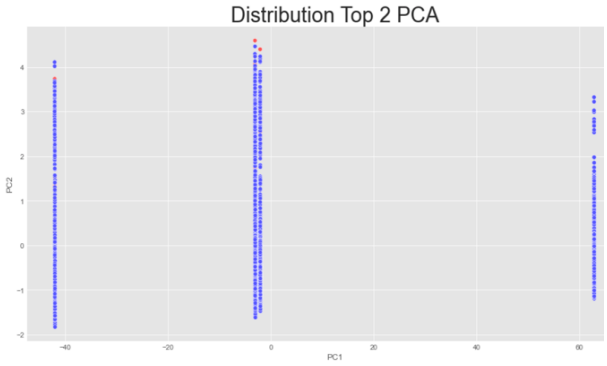
Kemudian untuk melihat persebaran data dapat juga menggunakan PCA (*Principal Component Analysis*). PCA adalah metode yang dipakai dalam *exploratory data analysis* untuk menvisualisasi variasi data serta mewakili semua variabel yang ada. Distribusi PCA seperti pada Gambar 11.



Gambar 9. BMI dan *avg_glucose_level*



Gambar 10. Persebaran data



Gambar 11. PCA

Tabel 6. Proporsi data

Data	Record	Variabel
Train set	4087	10
Test set	1022	10

Tabel 7. Proporsi data ROS

Sebelum		Sesudah	
1	**0**	**1**	**0**
198	3889	249	4861
198	3889	1549	3855

5.4 *Pre-Balancing* Data

Tahap ini bertujuan untuk memahami data dan pemilihan teknik serta metode *data mining*. *Pre-processing* yang berisi seleksi dan transformasi data, penghapusan data yang tidak dipakai, penambahan data yang hilang dan penghalusan data. Pada tahap ini juga berisi penerapan teknik *Data Balancing* menggunakan metode Random Forest. Dataset dibagi menjadi 2, yaitu untuk keperluan data training dan data testing. Presentase data dapat dilihat pada Tabel 6.

5.5 Penerapan *Data Balancing*

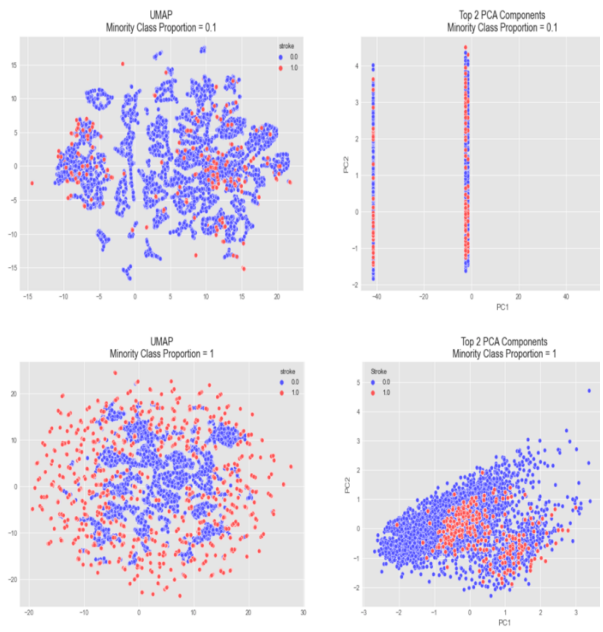
5.5.1 Teknik Random Oversampling

Teknik Random Oversampling (ROS) digunakan untuk membuat sintetis dari *class* minoritas dengan cara merandom data yang ada. Pada teknik ini digunakan *class* minoritas dengan proporsi 0,1 dan proporsi 1. Proporsi ini digunakan berdasarkan hasil yang didapatkan pada PCA. Tabel 7 menunjukan proporsi sintis yang sebelum dan sesudah penerapan teknik oversampling pada masing-masing *class* yang ada pada data *training*.

Gambar 12 menunjukkan persebaran antar data *class* mayoritas dan minoritas pada pada proporsi 0,1 dan 1 data *training*. Gambar 13 merupakan hasil yang didapatkan oleh metode Random Forest pada data training dan testing sebelum penerapan teknik *Data Balancing.*

5.5.2 Teknik Random Undersampling

Teknik Random Undersampling (RUS) digunakan untuk mengurangi class mayoritas dengan cara merandom data yang ada. Pada teknik ini menggunakan proporsi 0,1 dan 1. Proporsi ini digunakan berdasarkan hasil yang didapatkan pada proses PCA dan diperlakukan sama dengan ROS. Tabel 8 menunjukkan proporsi



Gambar 12. Persebaran data ROS

```
Training Classification_Report:
              precision    recall  f1-score   support

         0.0       0.95      1.00      0.98      3889
         1.0       1.00      0.04      0.08       198

    accuracy                           0.95      4087
   macro avg       0.98      0.52      0.53      4087
weighted avg       0.96      0.95      0.93      4087

Testing Classification_Report:
              precision    recall  f1-score   support

         0.0       0.95      1.00      0.97       971
         1.0       0.00      0.00      0.00        51

    accuracy                           0.95      1022
   macro avg       0.48      0.50      0.49      1022
weighted avg       0.90      0.95      0.93      1022
```

Gambar 13. Akurasi Data

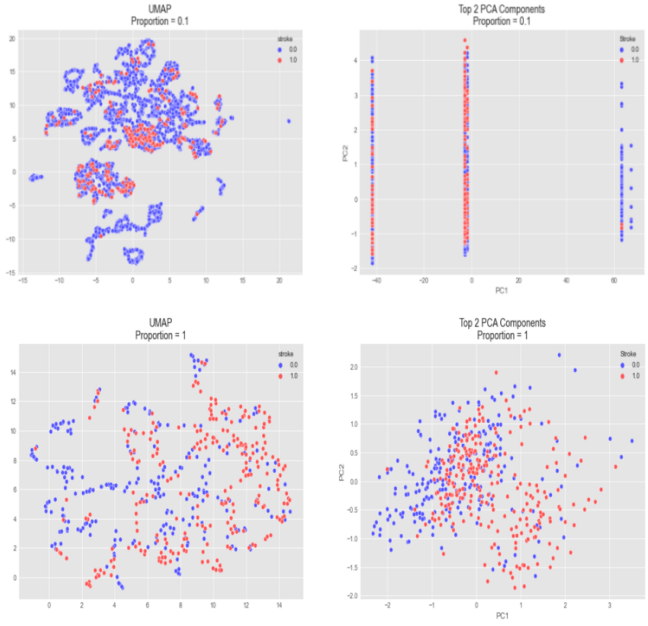
sintis yang sebelum dan sesudah penerapan teknik undersampling pada masing-masing *class* yang ada pada data *training*. Gambar 14 menunjukkan persebaran antar data *class* mayoritas dan minoritas pada pada proporsi 0,1 dan 1 data *training* yang menggunakan RUS.

Tabel 8. Persentase data RUS

Sebelum		Sesudah	
1	**0**	**1**	**0**
198	3889	249	4861
198	3889	1549	3855

Tabel 9. Prediksi Confusion Matriks

	PP	PN
Positive Class	True Positive	False Negative
Negative Class	False Positive	True Negative



Gambar 14. Persebaran data RUS

```
Testing Classification_Report:
              precision    recall  f1-score   support

         0.0       0.95      1.00      0.97       971
         1.0       0.00      0.00      0.00        51

    accuracy                           0.95      1022
   macro avg       0.48      0.50      0.49      1022
weighted avg       0.90      0.95      0.93      1022
```

Gambar 15. Evaluasi Pre-Balancing

```
Random Forest Results with Random Oversampling:
Proportion = 0.1
              precision    recall  f1-score   support

         0.0       0.96      0.94      0.95      4860
         1.0       0.16      0.21      0.18       249

    accuracy                           0.91      5109
   macro avg       0.56      0.58      0.56      5109
weighted avg       0.92      0.91      0.91      5109


Random Forest Results with Random Oversampling:
Proportion = 1
              precision    recall  f1-score   support

         0.0       0.95      1.00      0.98      4860
         1.0       0.00      0.00      0.00       249

    accuracy                           0.95      5109
   macro avg       0.48      0.50      0.49      5109
weighted avg       0.90      0.95      0.93      5109
```

Gambar 16. Evaluasi teknik ROS

5.6 Evaluasi Teknik

Evaluasi dilakukan menggunakan tabel Confusion Matriks. Tabel 9 digunakan untuk melihat hasil prediksi yang didapatkan setiap teknik dan mengukur performa yang didapatkan oleh metode dengan menggunakan teknik penanganan *imbalance data* [7].

$$\text{Akurasi} = \frac{f_{11}+f_{00}}{f_{11}+f_{10}+f_{01}+f_{00}} \quad (1)$$

5.6.1 Evaluasi Pre-Balancing Data

Evaluasi performa yang didapatkan oleh model Random Forest sebelum menerapkan *Data Balancing*. Gambar 15 menunjukkan tabel Confusion Matriks yang didapatkan oleh model tanpa *pre-balancing*.

5.6.2 Evaluasi Teknik Random Oversampling

Evaluasi performa dilakukan juga pada penerapan metode Random Forest dengan teknik ROS pada proporsi 0,1 dan 1. Gambar 16 adalah tabel Confusion Matriks yang dihasilkan dari teknik ROS.

5.6.3 Teknik Random Undersampling

Evaluasi performa dilakukan pada metode Random Forest pada penggunaan teknik ROS proporsi 0,1 dan 1. Gambar 17 adalah tabel Confusion Matriks yang dihasilkan dari teknik RUS. Perbandingan teknik ROS dan RUS dapat dilihat pada Tabel 10.

Tabel 10. Evaluasi Teknik

Proporsi	Sebelum	Sesudah	
		ROS	RUS
0.1	95%	91%	21%
1	95%	95%	76%

```
Random Forest Results with Random Undersampling:
Proportion = 0.1
              precision    recall  f1-score   support

         0.0       0.98      0.17      0.29      4860
         1.0       0.06      0.94      0.10       249

    accuracy                           0.21      5109
   macro avg       0.52      0.56      0.20      5109
weighted avg       0.94      0.21      0.28      5109

Random Forest Results with Random Undersampling:
Proportion = 1
              precision    recall  f1-score   support

         0.0       0.95      0.79      0.86      4860
         1.0       0.06      0.27      0.10       249

    accuracy                           0.76      5109
   macro avg       0.51      0.53      0.48      5109
weighted avg       0.91      0.76      0.83      5109
```

Gambar 17. Evaluasi teknik RUS

6. KESIMPULAN

Berdasarkan reviu literatur yang dilakukan, terdapat 2 (dua) faktor utama yang dapat mempengaruhi performa *machine learning* yaitu pemilihan metode dan teknik penanganan *imbalance data*. Jumlah dataset juga dapat mempengaruhi performa yang akan diberikan oleh model *machine learning.* Teknik yang tepat berkontribusi besar dalam klasifikasi dan mampu mendapatkan performa yang baik pada *class* mayoritas.

Pada klasifikasi kemungkinan penyakit stroke, akurasi metode Random Forest sebelum dan sesudah menerapkan teknik ROS adalah 95%. Penggunaan teknik ROS mendapatkan performa yang lebih tinggi yaitu 95% dari pada teknik RUS (76%). Pada penanganan *Imbalance Data* yang menggunakan teknik ROS dan tanpa penerapan *Data Balancing* memiliki performa atau akurasi yang sama. Namun dengan menerapkan *Data Balancing,* maka klasifikasi tidak akan condong pada *class* 0 (tidak stroke) dan mampu mengklasifikasikan *class* minoritas 1 (stroke) lebih baik.

7. REFERENSI

[1] W. Johnson, O. Onuma, M. Owolabi, & S. Sachdev. 2016. Stroke: A global response is needed, *Bull. World Health Organ.*, 94, 9, 634A-635A, doi: 10.2471/BLT.16. 181636.

[2] American Stroke Association. *About Stroke*. www.stroke.org. (Diakses 1 April 2021)

[3] P. Simbolon, N. Simbolon, & M. S. Ringo. 2018. Faktor Merokok dengan Kejadian Stroke di Rumah Sakit Santa Elisabeth Medan. *Jurnal Kesehatan Manarang.* 4, 1, 18, doi: 10.33490/jkm.v4i1.53.

[4] Kementerian Kesehatan Republik Indonesia. 2012. *Penyakit Tidak Menular.* Publ. DATA DAN INFORMASI, Buletin. [Online] Available: http://www.depkes.go.id/download.php?file=download/pusdatin/buletin/buletin-ptm.pdf.

[5] Suyanto, 2019. *Data Mining untuk Klasifikasi dan Klasterisasi Data Edisi Revisi.* Bandung: Informatika.

[6] S. Fotouhi, S. Asadi, & M. W. Kattan. 2019. A comprehensive data level analysis for cancer diagnosis on imbalanced data. *Journal of Biomedical Informatics.* 90, February, 103089. doi: 10.1016/j.jbi.2018. 12.003.

[7] Hairani, N. A. Setiawan, & T. B. Adji. 2013. Metode Klasifikasi Data Mining dan Teknik Sampling Smote. *Seminar Nasional Sains dan Teknologi*, 168–172.

[8] A. A. Oscar Febri Ramadhan & Adiwijaya. 2017. *Handling Imbalanced Data pada Prediksi Churn menggunakan metode SMOTE dan KNN Based on Kernel*. Skripsi. Telkom University. 4, 117, 1–15.

[9] A. N. Kasanah, Muladi, & U. Pujianto. 2019. Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam, *RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 3, 10.

[10] H. Marcos & H. Utomo. 2015. Perbandingan Kinerja Algoritme C.45 Dan Naive Bayes Mengklasifikasi Penyakit Diabetes, *Jurnal Informatika Darmajaya.* 15, 2, 141–148. doi: 10.30873/ji.v15i2.596.

[11] U. R. Salunkhe & S. N. Mali. 2016. Classifier Ensemble Design for Imbalanced Data Classification: A Hybrid Approach, *Procedia Computer Science.* 85, 725–732, doi: 10.1016/j.procs.2016.05.259.

[12] A. Syukron & A. Subekti. 2018. Penerapan Metode Random Over-Under Sampling dan Random Forest Untuk Klasifikasi Penilaian Kredit. *Jurnal Informatika.* 5, \2, 175–185. doi: 10.31311/ji.v5i2.4158.

[13] M. Hayaty, S. Muthmainah, & S. M. Ghufran. 2020. Random and Synthetic Over-Sampling Approach to Resolve Data Imbalance in Classification. *International Journal Of Artificial Intelligence Research.* 4, 2, 86. doi: 10.29099/ijair.v4i2.152.

[14] Asniar, N. U. Maulidevi, & K. Surendro. 2021. SMOTE-LOF for noise identification in imbalanced data classification. *Journal of King Saud University - Computer and Information Sciences.* In Press. doi: 10.1016/j.jksuci.2021.01.014.