Jutisi: Jurnal Ilmiah Teknik Informatika dan Sistem Informasi
Jl. Ahmad Yani, K.M. 33,5 - Kampus STMIK Banjarbaru
Loktabat – Banjarbaru (Tlp. 0511 4782881), e-mail: puslit.stmikbjb@gmail.com
e-ISSN: 2685-0893
p-ISSN: 2089-3787

Penerapan Metode *Smote Extreme Gradient Boosting* Untuk Klasifikasi Penyakit Kanker *Serviks* Di Kota Medan

Debi Anggitasyah[1*], Machrani Adi Putri Siregar[2]
Matematika, Universitas Islam Negeri Sumatera Utara, Medan, Indonesia
*e-mail *Corresponding Author*: debi0703192001@uinsu.ac.id

Abstract

Cervical cancer or often referred to as cervical cancer is a cancer that forms cervical tissue, cases of cervical cancer occur every year, making it the number 2 killer in Indonesia. This is very concerning considering that cervical cancer is one of the cancers that can be prevented early on but is very difficult to detect. Therefore, a classification is needed for prevention detection, the problem in classification is the classification of unbalanced data where the distribution of samples in all known classes of observations does not have the same proportions. The purpose of this study was to determine the application and results of the accuracy of the Smote Extreme Gradient Boosting method in classifying cervical cancer in hospitals. Dr. Pirngadi, Medan City. This research is a quantitative research, using the smote extreme gradient method. This study found 69 cases of cervical cancer classified using the XGBOOST algorithm and the Smote algorithm as a data imbalance problem solver which is a very good solution because the results of the accuracy of the Area Under Cover (AUC) Smote extreme gradient boosting method yield a value of 1.00% which is classified as the best classification.

Keywords*: Binary Classification; SmoteXgboot; Area Under Cover; Matrix; Variables*

Abstrak

Kanker serviks atau sering disebut dengan kanker leher rahim merupakan kanker yang membentuk jaringan leher rahim, kasus kanker serviks terjadi setiap tahunnya, menjadikannya pembunuh nomor 2 di Indonesia. Hal ini sangat memprihatinkan mengingat kanker serviks merupakan salah satu kanker yang dapat dicegah sejak dini tetapi sangat sulit dideteksi keberadaannya. Oleh karena itu diperlukan klasifikasi untuk deteksi pencegahannya, permasalahan dalam klasifikasi adalah klasifikasi data yang tidak seimbang dimana sebaran sampel pada semua kelas pengamatan yang diketahui tidak memiliki proporsi yang sama. Tujuan penelitian ini adalah untuk mengetahui penerapan dan hasil keakuratan metode *Smote Extreme Gradient Boosting* dalam mengklasifikasi penyakit kanker serviks di RSUD. Dr. Pirngadi Kota Medan. Penelitian ini merupakan penelitian kuantitatif, menggunakan metode *smote extreme gradient.* Penelitian ini didapatkan 69 kasus kanker serviks yang diklasifikasikan menggunakan algoritme *XGBOOST* dan algoritme *Smote* sebagai pemecah masalah ketidakseimbangan data sangat menjadi solusi karena pada hasil tingkat keakuratan *Area Under Cover* (AUC) metode *Smote extreme gradient boosting* menghasilkan nilai 1.00% yang mana klasifikasi tergolong dalam klasifikasi terbaik.

Kata kunci: *Klasifikasi Biner; SmoteXgboot; Area Under Cover; Matrix; Variabel*

1. Pendahuluan

Kanker serviks adalah kanker yang tumbuh dan berkembang di leher rahim atau serviks, terutama dari lapisan epitel atau lapisan terluar dari permukaan serviks. Serviks (leher rahim) adalah bagian rahim atau rahim yang berada di 1/3 bagian bawah rahim [1]. Serviks mengandung sel epitel yang sangat rentan terhadap masuknya kuman penyakit. Di bagian inilah virus penyebab kanker tumbuh. Penyebab kanker serviks berawal dari sel normal serviks yang terinfeksi virus Human Papilloma Virus. HPV adalah virus DNA yang berukuran 8.000 pasangan basa berbentuk ikosahedral dengan ukuran 55nm, memiliki 72 kapsomer dan 2 protein kapsid. Karena ukuran virus HPV sangat kecil, maka virus ini dapat menular melalui lesi mikro atau sel abnormal pada vagina [2]. Jenis virus HPV yang paling berisiko adalah HPV 16 dan 18 yang

sering ditemukan pada kanker dan lesi prakanker serviks yang menyebabkan kerusakan sel mukus bagian luar yang berujung pada keganasan [3]. Kanker ini sekelompok penyakit yang ditandai dengan pertumbuhan dan penyebaran sel abnormal yang tidak terkendali [4].

Klasifikasi adalah proses untuk mendapatkan rumus fungsi yang mewakili label kelas tertentu dari data tertentu untuk prediksi lebih lanjut kelas yang belum diketahui sebelumnya [5]. Klasifikasi adalah suatu bentuk analisis data dengan menyajikan model data penting melalui fitur tertentu [6][7]. *Preprocessing* data bertujuan untuk mengubah data input mentah menjadi format yang sesuai untuk analisis lebih lanjut [8]. Salah satu masalah umum yang diselesaikan pada tahap preprocessing data adalah menangani nilai yang hilang *(missing value)*. *Missing value* terjadi ketika ada beberapa informasi yang tidak tersedia untuk subjek (kasus) yang biasanya karena kesalahan input data, informasi tentang subjek tidak tersedia atau tidak tersedia. *Missing value* bisa diatasi dengan mengisi data yang hilang, maka data yang hilang diperhitungkan dengan nilai rata-rata jika datanya kuantitatif dan modusnya jika datanya kualitatif [9]. *Syntethic Minority Oversampling Technique (SMOTE)* adalah metode populer untuk mengatasi data yang tidak seimbang, pada generating samples misalkan 3 sampel digunakan sebagai sampel acuan dan akan menghasilkan sampel baru kemudian dilakukan berulang, maka sampel baru yang muncul akan melengkapi dataset minoritas yang sedikit, sehingga akhirnya dataset berwana biru dan hijau akan menjadi seimbang di resampled dataset.

Gambar 1. Cara Kerja *SMOTE*

Penelitian ini dilakukan bertujuan untuk mengetahui penerapan metode *Smote Extreme Gradient boosting* dalam mengklasifikasikan faktor penyebab penyakit kanker serviks di kota Medan dan mengetahui hasil keakuratan metode *Smote Extreme Gradient Boosting* dalam mengklasifikasikan faktor penyebab penyakit kanker serviks di kota Medan. Hasil penelitian ini diharapkan dapat diperoleh informasi mengenai variabel apa saja yang berpengaruh secara signifikan terhadap penyebab faktor-faktor kanker serviks dari penelitian ini berguna sebagai upaya kecil membantu pemerintah dan pemangku kepentingan lainnya dalam membuat kebijakan dan tindakan yang tepat untuk memperbaiki masalah ini sebagaimana mestinya juga untuk kehidupan masyarakat dikota medan yang lebih baik lagi.

2. Tinjauan Pustaka

Penelitian sebelumnya yang membahas tentang metode XGBOOST oleh Pinata dkk pada tahun 2020 mengenai Prediksi Kecelakaan Lalu Lintas di Bali dengan *Xgboost* pada *Python*. Penelitian ini dilakukan untuk meramalkan jumlah kejadian, jumlah orang meninggal dunia, jumlah orang yang mengalami luka ringan, dan luka berat pada setiap tahunnya menghasilkan nilai error yang cukup rendah [10].

Penelitian Elok pada tahun 2021 mengenai Klasifikasi Penyakit *Stroke* Menggunakan Metode *Smote Xgboost*. Penelitian ini dilakukan untuk membandingkan kinerja dan mencari hasil *area under curve* antara metode *xgboost* dengan metode *smote xgboost* [11].

Penelitian selanjutnya yang membahas tentang metode XGBOOST oleh Ubaidillah pada tahun 2022 mengenai Implementasi *Xgboost* Pada Keseimbangan Liver Patient Dataset dengan SMOTE dan *Hyperparameter Tuning Bayesian Search*. Penelitian ini dilakukan untuk menangani masalah ketidakseimbangan kelas pada Indian Liver Pattient Dataset. Berbeda dengan. Peningkatan performa model Smote-Xgboost menggunakan *hyperparameter tuning Bayesian Search* untuk mendapatkan model yang optimal [12].

Penelitian ini penting dilakukan karena untuk mengetahui penerapan metode *Smote Extreme Gradient Boosting* dalam melihat dari faktor penyebab mana yang paling banyak terjangkit kanker serviks serta untuk mengetahui hasil tingkat keakuratan metode *Smote Extreme Gradient Boosting* dalam mengklasifikasi penyakit kanker serviks. Penelitian ini dilakukan hanya untuk menangani masalah ketidakseimbangan kelas pada *Indian Liver Pattient Dataset*. Adapun

fokus penelitian ini adalah Data yang digunakan adalah data penyakit kanker serviks tahun 2018 sampai dengan tahun 2022 yang diambil dari RSUD Dr. Pirngadi Kota Medan. Masalah klasifikasi pada penelitian ini adalah masalah klasifikasi biner, yaitu terdapat dua jenis klasifikasi (0 dan 1). Proses analisis menggunakan aplikasi *Python*.

3. Metodologi

Extreme Gradient Boosting (XGBoost) pertama kali diusulkan oleh *Chen* dan *Guestrin* pada tahun 2016 untuk mengatasi berbagai masalah pembelajaran yang dapat menghasilkan hasil yang efisien, cepat, dan terukur dengan menerapkan *Gradient Tree Boosting* [13]. *Gradient Boosting* menggunakan ensambel dari pohon keputusan (*decision tree*) untuk memprediksi nilai. *Gradient boosting* membuat pohon keputusan berkelanjutan yang sederhana di mana setiap pohon baru yang dibuat merupakan peningkatan dari pohon kesalahan sebelumnya. Metode *XGBoost* menggunakan data training dengan beberapa fitur x_i untuk memprediksi variabel target y_i. Nilai yang diprediksi dapat memiliki interpretasi yang berbeda, tergantung pada tugasnya, yaitu regresi atau klasifikasi [14].

Pada kasus data penelitian yang tidak seimbang menggunakan algoritme *XGBoost*, digunakan parameter *scale_pos_weight*. Nilai parameter *scale_pos_weight* digunakan untuk mengoreksi kesalahan yang dihasilkan oleh model selama pelatihan pada data kelas positif. Sehingga model memiliki kinerja yang baik dalam memprediksi kelas positif. Nilai parameter scale_pos_weight dihitung dari jumlah total kelas mayoritas dibagi dengan jumlah kelas minoritas pada data penelitian [15].

Hyperparameter Tuning digunakan untuk menentukan kombinasi nilai parameter yang tepat untuk memaksimalkan kinerja pemodelan, Salah satu *hyperparameter tuning* yang dapat digunakan adalah *grid search*. *Grid search* merupakan metode alternatif yang digunakan untuk mencari parameter terbaik dalam model dengan mencoba satu kombinasi parameter dalam satu waktu dan memvalidasi setiap kombinasi.

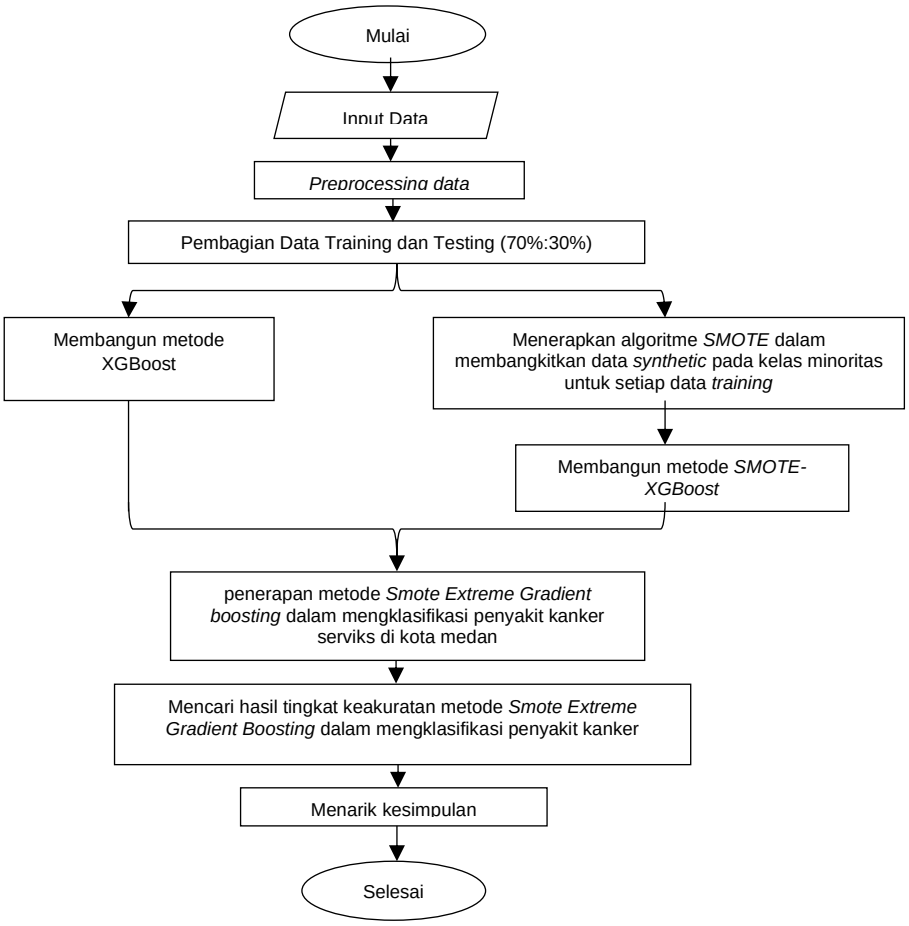
Area Under Curve (AUC) adalah area di bawah kurva untuk mencari hasil tingkat keakuratan dari satu ukuran kinerja klasifikasi dan untuk mengevaluasi model mana yang rata-rata lebih baik. Nilai *AUC* diperoleh dengan menghitung true positive rate (TPR) yaitu objek pada kelas positif yang diklasifikasikan dengan benar dan false positive rate (FPR) yaitu banyaknya objek pada kelas positif yang salah diklasifikasikan [16].

Penelitian ini merupakan jenis penelitian kuantitatif dan Metode penelitian yang digunakan adalah metode analisis data sekunder. yang bertujuan untuk menerapkan metode *Smote Extreme Gradient Boosting* dan mencari hasil keakuratan metode *Smote Extreme Gradient Boosting* dalam klasifikasi penyakit kanker serviks di kota medan. Teknik pengambilan sampel dalam penelitian ini adalah total sampling. Total sampling adalah teknik pengambilan sampel dimana jumlah sampel sama dengan jumlah populasi [17]. Alasan pengambilan total sampling karena jumlah populasi kurang dari 100 maka seluruh populasi digunakan sebagai sampel penelitian. Sampel yang diambil dari penelitian ini adalah 69 orang. Analisis data yang dilakukan dalam penelitian ini adalah analisis klasifikasi biner (*decision tree*) dengan menggunakan metode *Smote Extreme Gradient Boosting* dalam mengklasifikasi penyakit kanker serviks di kota medan [18].Penelitian ini menggunakan bantuan *software* python dengan *jupyterlab*. Prosedur penelitian yang akan dilakukan dalam penelitian ini sebagai berikut :

1) Memasukkan data penelitian untuk pemodelan dalam perangkat lunak Python.
2) Lakukan preprocessing data dengan mendeteksi keberadaan missing value. Jika tidak ada missing value maka akan dilanjutkan dengan pendeteksian *outlier*.
3) Membagi data penelitian menjadi data training dan data testing dengan perbandingan 70%-30%
4) Menerapkan hyperparameter algoritme *Smote-XGBoost* ke data penelitian menggunakan parameter *leaming_rate, max_depth, dan min_child_weight* [19].
   Langkah-langkahnya adalah sebagai berikut:
   a. Membuat prediksi awal atau base_score = 0.5 untuk semua data penelitian [20].
   b. Menghitung residual ẏ pada semua data dari prediksi sebelumnya.
   c. Membangun pohon dengan membagi data menjadi dua bagian dari berbagai kemungkinan pemisahan.
   d. Menghitung *similarity* (kesamaan) dan mendapatkan nilai untuk semua pohon yang dibangun untuk menemukan pohon dengan pemisahan optimal berdasarkan rumus.

e. Melakukan split lagi untuk pohon yang memiliki nilai gain maksimum sehingga *max_depth* konstruksi pohon selesai.
f. Melakukan pemangkasan atau pruning untuk memperkecil ukuran pohon dengan membuang bagian pohon yang memiliki nilai selisih gain dan gamma < 0.
g. Menghitung nilai *output* untuk semua leaf pohon.
h. Menghitung prediksi dari model yang terbentuk.
i. Membentuk *confusion matrix* dan hitung kinerja metode klasifikasi menggunakan nilai akurasi area di bawah kurva (AUC).
j. Menarik kesimpulan [21][22].

5) Menerapkan algoritme *SMOTE-XGBoost* pada data penelitian, langkah-langkahnya adalah sebagai berikut:
   a. Menentukan kelas data minoritas.
   b. menentukan nilai tetangga dengan k=5.
   c. Menghitung jarak antar data kelas minoritas dengan metode Euclidean.
   d. Melakukan perhitungan untuk menghasilkan data sintesis [23].
6) Menerapkan hyperparameter tuning dari algoritme *SMOTE-XGBoost* untuk meneliti data.
7) Menentukan hasil tingkat keakuratan metode *Smote Extreme Gradient Boosting* dalam mengklasifikasi faktor penyebab kanker serviks di kota Medan menggunakan nilai tingkat akurasi AUC [24][25][26].



Gambar 2. Diagram Alur Penelitian (*flowchart*)

Data penelitian yang digunakan adalah data sekunder yang berdasarkan data rekam medis yang ada di RSUD Dr. Pirngadi Kota Medan dengan jumlah 69 pasien, seluruh data yang digunakan adalah data pada tahun 2018 sampai 2022. Penyakit kanker serviks pasien dianalisis menggunakan data faktor-faktor yang mempengaruhinya, data yang digunakan terdiri dari 7 variabel independen dan 1 variabel. Berikut merupakan variabel yang digunakan dalam penelitian ini:

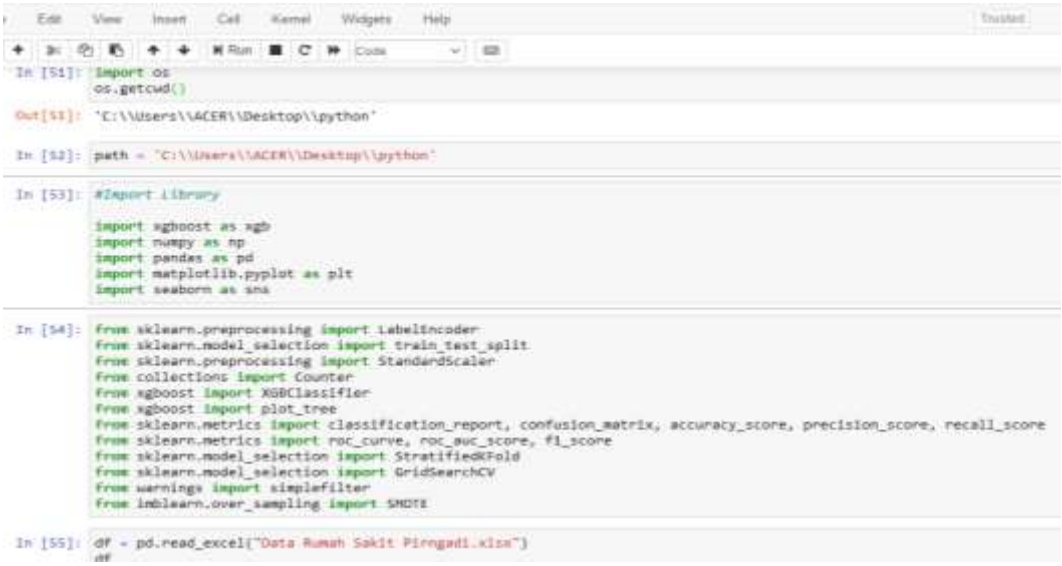
Tabel 1. Variabel Penelitian

No.	Variabel	Deskripsi	Simbol	Tipe Data
1	Y	Kanker serviks	Diagnosa	Kategori
2	X_1	Umur	Age	Numerik
3	X_2	penggunaan pembalut	Softex	Numerik
4	X_3	Penggunaan KB Spiral	KBSpiral	Kategori
5	X_4	Contraceptives	Contraceptives	Numerik
6	X_5	Pregnancies	Pregnancies	Numerik
7	X_6	smokes	Smokes	Kategori
8	X_7	Riwayat Keputihan	Vaginaldischarge	Kategori

4. Hasil dan Pembahasan

4.1 Deskripsi Data

Sebelum melakukan analisis data ditampilkan terlebih dahulu statistik deskriptif dari variabel penelitian yang digunakan. Statistik deskriptif ini berguna untuk menampilkan gambaran umum tentang data yang digunakan, total terdapat 69 data yang dibagi menjadi 7 variabel independen dan 1 variabel dependen. Berikut merupakan statistik deskriptif data yang dilakukan dengan bantuan *software python.*



Gambar 3. Input data *python jupyterlab*

Dengan bantuan *sofware python* berikut ditampilkan statistik deskriptif data yang bertujuan untuk memberikan gambaran umum dan menginterpretasikan data seperti nilai minimum, maximum, mean, dan srandart deviasi sebagai langkah awal dalam analisis statistik lebih lanjut serta data yang ditampilkan sudah menggunakan skala:

Tabel 2. Statistik Deskriptif data

	X1_ Age	X2_ Softex	X3_ KBSpiral	X4_ Contraceptives	X5_ Pregnancies	X6_ Smokes	X7_ Vaginaldischarge
Jumlah data	69	69	69	69	69	69	69
Mean	49,68	2,43	0,59	5,05	3,07	0,82	0,17
Standart deviasi	9,92	0,71	0.49	4,70	1,43	0,38	0,38
Min	28	1	0	0	0	0	0
Max	69	4	1	33	8	1	1

Selanjutnya memasukkan sintax untuk menampilkan plot diagram variabel tersebut dan menampilkan hasil persentase akhir dari plot diagram variabel diatas:

Gambar 4. Kode Untuk Plot Diagram

Tabel 3. Hasil Ukur Skala Menggunakan *SmoteXGBoost*

Age	Softex	KBSpiral	Contraceptives	Pregnancies	Smokes	Vaginal discharge
<40 = 0.20	<2 = 0.63	Tidak = 0.40	<4 = 0.37	<3 = 0.40	Tidak = 0.82	Tidak = 0.17
>40 = 0.79	>2 = 0.36	Ya = 0.59	>4 = 0.62	>3= 0.59	Ya = 0.17	Ya = 0.82

Dapat dilihat pada tabel 3 bahwa pada variabel Age, usia lebih dari 40 tahun dengan 79% lebih besar resikonya terserang penyakit kanker serviks, pengggantian pembalut kurang dari 2 kali dalam sehari dengan 63% lebih besar resikonya terserang penyakit kanker serviks, penggunaan alat Kb spiral dengan 59% lebih besar resikonya terserang penyakit kanker serviks daripada yang tidak menggunakan, penggunaan alat kontrasepsi lebih dari 4 kali dengan 62% lebih besar resikonya terserang penyakit kanker serviks, memiliki jumlah anak lebih dari 3 dengan 59% lebih besar resikonya terserang penyakit kanker serviks, riwayat merokok pada tabel diatas disimpulkan bahwa hanya berdampak 17% resikonya terserang penyakit kanker serviks dan terakhir riwayat keputihan pada wanita dengan 82% lebih besar resikonya terserang penyakit kanker serviks.

4.2 Preprocessing Data

Melakukan preprocessing data dengan mendeteksi keberadaan *missing value*. Periksa adanya *missing value*, jika data yang digunakan dalam penelitian terdapat *missing value* maka akan dilakukan imputasi dengan menggunakan nilai rata-rata (variabel kontinu) atau menggunakan nilai modus (variabel diskrit). Jika tidak ada missing value maka akan dilanjutkan dengan pemangkasan *(drop)* kolom tabel yang selain dari 7 variabel independen dan 1 variabel dependen dengan menggunakan *software python*.

Gambar 5. Kode Untuk *Missing Value* dan *Drop*

Dengan menggunakan fungsi isna() dan sum() pada *missing value* kita tahu bahwa dalam dataset semua kolom tidak ada nilai yang kosong. Jadi fungsi keduanya adalah untuk mengisi nilai yang kosong (hilang).

```
Out[67]: No_Column              0
         No_RekamMedis          0
         Y_Diagnosa             0
         Year                   0
         X1_Age                 0
         X2_Softex              0
         X3_KBSpiral            0
         X4_Contraceptives      0
         X5_Pregnancies         0
         X6_Smokes              0
         X7_Vaginaldischarge    0
         dtype: int64
```

Gambar 6. Output *Missing value*

Diketahui bahwa pada dataset *missing value* bernilai 0 itu tandanya tidak ada nilai angka yang hilang atau kosong. Selanjutnya dengan menggunakan kode *label encoder* untuk mengubah data kategori menjadi data numerik.

```
In [330]: #Mengubah Data Kategori Menjadi Numerik

label_encoder=LabelEncoder()
df['X1_Age']=label_encoder.fit_transform(df['X1_Age'])
df['X2_Softex']=label_encoder.fit_transform(df['X2_Softex'])
df['X3_KBSpiral']=label_encoder.fit_transform(df['X3_KBSpiral'])
df['X4_Contraceptives']=label_encoder.fit_transform(df['X4_Contraceptives'])
df['X5_Pregnancies']=label_encoder.fit_transform(df['X5_Pregnancies'])
df['X6_Smokes']=label_encoder.fit_transform(df['X6_Smokes'])
df['X7_Vaginaldischarge']=label_encoder.fit_transform(df['X7_Vaginaldischarge'])
df
```

Out[330]:

	X1_Age	X2_Softex	X3_KBSpiral	X4_Contraceptives	X5_Pregnancies	X6_Smokes	X7_Vaginaldischarge	Y_Diagnosa
0	1	1	0	0	0	1	0	Kanker Serviks
1	1	0	1	1	1	1	0	Kanker Serviks
2	0	0	1	0	1	1	0	Kanker Serviks
3	1	0	1	1	1	1	0	Kanker Serviks

Gambar 7. Kode untuk Mengganti Data Kategori Menjadi Data Numerik

4.3 Pembagian *Training* dan *Testing*

Selanjutkan dilakukan Membagi dataset menjadi data *training* dan data *testing* dengan perbandingan 70%-30%. Sebelum masuk ke tahap analisis klasifikasi. Data *training* digunakan untuk melatih algoritme dalam pembentukan model sedangkan data *testing* digunakan untuk menilai performa model yang didapatkan dari data *training*.

```
In [18]: #Features And Target (X,y)

X = df.drop('Y_Diagnosa', axis=1)
y = df['Y_Diagnosa']
print(X.shape)
print(y.shape)

X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.3,random_state=42)
print("Jumlah X_train dataset: ",X_train.shape)
print("Jumlah y_train dataset: ",y_train.shape)
print("Jumlah X_test dataset: ",X_test.shape)
print("Jumlah y_test dataset: ",y_test.shape)

#Feature Scaling

standard_scaler = StandardScaler()
X_train=standard_scaler.fit_transform(X_train)
X_test=standard_scaler.fit_transform(X_test)
print(X_train)
print(X_test)

X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.3,random_state=42)
X_train.shape, X_test.shape
```

Gambar 8. Kode Untuk *Training* dan *Testing*

Dengan menggunakan fungsi *drop* untuk menghapus beberapa kolom yang tidak ingin dihitung kemudian menggunakan fungsi *feature scaling* dengan *standart scaler* untuk memproses pembagian data *training* dan data *testing*

Tabel 4. Proporsi data *Training* dan *Testing*

	Training	Testing	Total
Rasio	70%	30%	100%
Jumlah	48	21	69

Diketahui bahwa data penelitian sebanyak 69 data terbagi menjadi 48 data *training* dan 21 *data testing*. Dalam kasus ini, dari 48 data training diketahui bahwa kelas 0 berjumlah 40 data sedangkan kelas 1 hanya berjumlah 8 data. Sehingga dapat dikatakan bahwa data penelitian ini tidak seimbang karena kelas data pada data training tidak memiliki proporsi yang sama.

4.4 *Hyperparameter Tuning*

Tahapan selanjutnya yaitu melakukan *tuning hyperparameter Smote XGBoost* ke data penelitian menggunakan parameter *leaming_rate, max_depth, dan min_child_weight*, Membuat prediksi awal atau *base_score = 0.5* Menghitung residual ẏ, Menghitung *similarity* (kesamaan). Melakukan *split*. Melakukan pemangkasan atau *pruning*.

```
In [181]: #tuning hyperparameter SMOTEXGBOOST (grid search)

          model1=XGBClassifier(objective='binary:logistic',scale_pos_weight=1)
          max_depth=[4,5,6,7,8]
          min_child_weight=[1,2,3,4,5,6]
          learning_rate=[0.01,0.05,0.1,0.3,0.5]

          param_grid=dict(max_depth=max_depth, min_child_weight=min_child_weight,learning_rate=learning_rate)
          kfold=StratifiedKFold(n_splits=5, shuffle=True, random_state=0)
          grid_search=GridSearchCV(model1, param_grid, scoring="accuracy", n_jobs=4, cv=kfold)
```

```
In [182]: #Fit XGBoost model on training data
          model1=XGBClassifier(objective='binary:logistic',max_depth=6,min_child_weight=1,learning_rate=0.3,n_jobs=4)
          model1.fit(X_train, y_train)
```

Gambar 9. Kode Untuk *Hyperparameter Tuning*

Dengan menggunakan fungsi *Grid Seacrh* untuk mencari *tuning hyperparameter Smote XGBoost* untuk mengklasifikasikan data penelitian. Parameter *tuning* yang digunakan adalah *learning_rate, max_depth dan min_child_weight*. Serta digunakan parameter *scale pos_weight=1* untuk mengatasi ketidakseimbangan pada data penelitian.

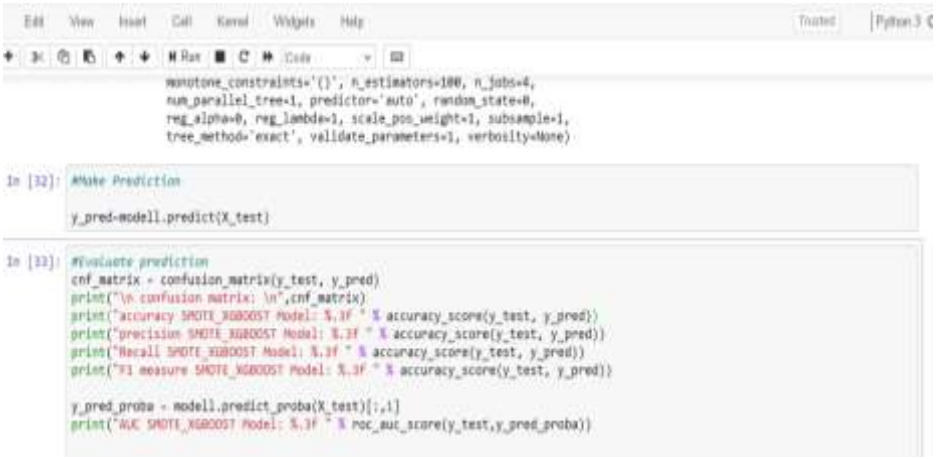
Tabel 5. Hasil *Tuning Hyperparameter* Metode *Smote Xgboost*

	Nilai parameter	Parameter terbaik
Learning_rate	0.001, 0.01, 0.1, 0.3	0.3
Max_depth	2,3,4,5,6,7	6
min_child_weight	1,2,3,4	1

Berdasarkan hasil *tuning hyperparameter XGBoost* yang ditampilkan oleh tabel 5. menunjukkan bahwa dari nilai parameter *Learning_rate* 0.001, 0.01, 0.1 diperoleh nilai parameter *Learning_rate* tebaik yaitu 0.3, dari nilai parameter *Max_depth* 2,3,4,5,6,7diperoleh nilai parameter *Max_depth* terbaik yaitu 6 dan dari nilai parameter *min_child_weight* 1,2,3,4, diperoleh nilai parameter *min_child_weight* terbaik yaitu 1. Kemudian nilai parameter terbaik ini digunakan dalam membuat model klasifikasi.

4.5 Confusion Matrix

Menerapkan algoritme *SMOTE-XGBoost* pada data penelitian, langkah berikutnya melakukan evaluasi kinerja model dengan melakukan prediksi terhadap data *testing* menggunakan model pohon *XGBoost***,** menghitung prediksi dari model yang terbentuk. Membentuk *confusion matrix* dan hitung kinerja metode klasifikasi menggunakan nilai akurasi area di bawah kurva *(AUC).*



```
monotone_constraints='()', n_estimators=100, n_jobs=4,
num_parallel_tree=1, predictor='auto', random_state=0,
reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=1,
tree_method='exact', validate_parameters=1, verbosity=None)

In [32]: #Make Prediction

         y_pred=modell.predict(X_test)

In [33]: #Evaluate prediction
         cnf_matrix = confusion_matrix(y_test, y_pred)
         print("\n confusion matrix: \n",cnf_matrix)
         print("accuracy SMOTE_XGBOOST Model: %.3f " % accuracy_score(y_test, y_pred))
         print("precision SMOTE_XGBOOST Model: %.3f " % accuracy_score(y_test, y_pred))
         print("Recall SMOTE_XGBOOST Model: %.3f " % accuracy_score(y_test, y_pred))
         print("F1 measure SMOTE_XGBOOST Model: %.3f " % accuracy_score(y_test, y_pred))

         y_pred_proba = modell.predict_proba(X_test)[:,1]
         print("AUC SMOTE_XGBOOST Model: %.3f " % roc_auc_score(y_test,y_pred_proba))
```

Gambar 10. Kode Untuk prediksi *Confusion Matrix*

Dengan menggunkan fungsi *predict* untuk memprediksi data testing dan juga untuk menampilkan hasil *Confusion Matrix* dengan menggunakan fungsi *predict_proba.*

Tabel 6. *Confusion Matrix* Metode *SmoteXgboost*

Aktual	**Prediksi**	
	Positif	**Negatif**
Positif	17	0
Negatif	0	4

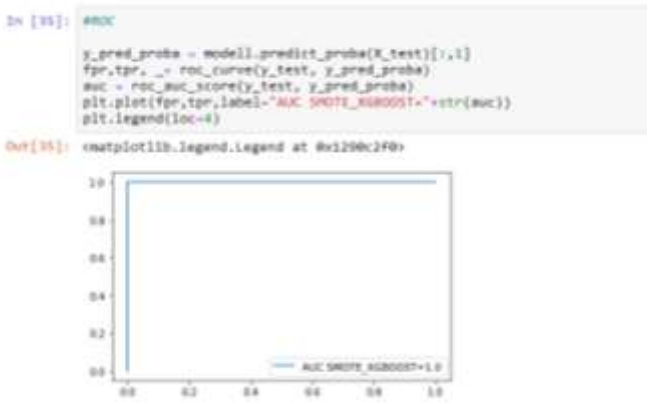
Tabel 6 menunjukkan bahwa model klasifikasi *SmoteXGBoost* menggunakan parameter *Learning_rate* =0.3, *Max_depth*=6, *min_child_weight*=1, memiliki ketepatan klasifikasi memprediksi penyakit kanker serviks dengan prediksi benar sebanyak 15 dan yang tidak tepat dalam memprediksi penyakit stroke pasien sebanyak 6 dengan jumlah data 69. Dari tabel *confusion matrix* didapatkan matrix pengujian menggunakan metode *XGBoost* yang ditunjukkan sebagai berikut:

Tabel 7. Metric pengujian Metode *SmoteXGBoost*

Metric pengujian	**Nilai**
Akurasi	1.00
Presisi	0.97
Recall	0.92
F1-Measure	0.82
AUC	1.00

Tabel 7 menunjukkan bahwa model *SmoteXgboost* dapat mengklasifikasikan penyakit kanker serviks dengan nilai akurasi, presisi, recall, AUC atau ketepatan model dalam memprediksi data.

4.6 Hasil Tingkat Keakuratan *SmoteXGBoost*



Gambar 11. Kode Untuk Keakuratan AUC

Berdasarkan gambar 11 menunjukkan bahwa metode *SmoteXgboost* menghasilkan nilai AUC sebesar 1.00% maka dapat disimpulkan bahwa apabila dari 69 jumlah pemeriksaan pasien yang menghasilkan kesimpulan besar dalam menentukan terjadinya kanker serviks dengan gejala faktor yang ada dan nilai AUC ini termasuk nilai klasifikasi sangat baik.

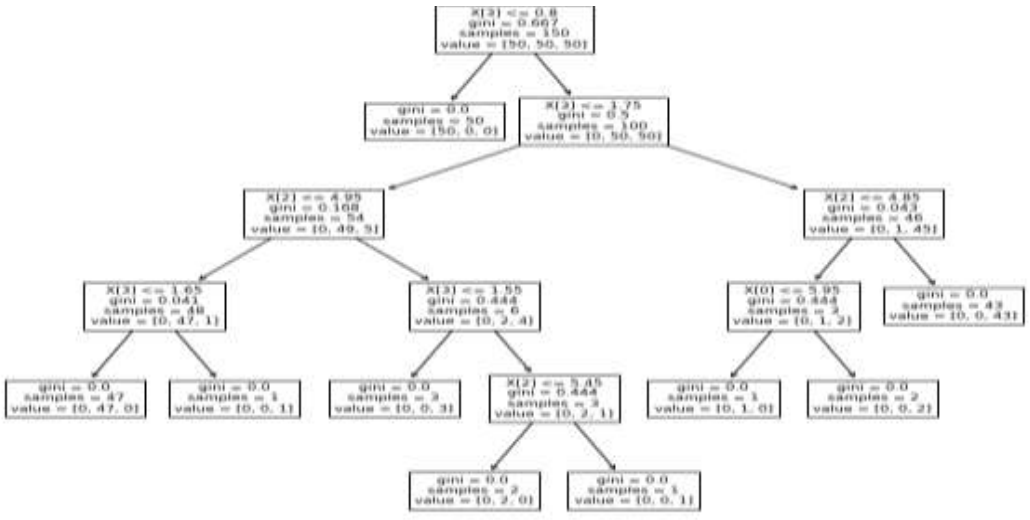
Tabel 8 Metric pengujian Metode *SmoteXGBoost*

Kelas	Sebelum	Sesudah
0	40	40
1	8	40
Jumlah	48	80

Tabel 8 menunjukkan bahwa kelas 0 dan 1 sudah bejumlah sama yaitu masing-masing berjumlah 40 data. Sehingga dapat dikatakan bahwa masalah ketidakseimbangan data sudah teratasi menggunakan algoritme Smote.

Dapat disimpulkan bahwa pada kasus penyakit kanker serviks yang diklasifikasikan menggunakan algoritme XGBOOST dan algoritme Smote sebagai pemecah masalah ketidakseimbangan data sangat menjadi solusi karena pada hasil tingkat keakuratan AUC metode Smote extreme gradient boosting menghasilkan nilai 1.00% yang mana klasifikasi tergolong dalam klasifikasi terbaik.

Pohon Klasifikasi Biner Penyakit Kanker Serviks:



Gambar 12. Pohon Klasifikasi Biner

4.7 Pembahasan

Berdasarkan 69 data yang ada di RSUD Dr. Pirngadi Kota Medan, seluruh data yang digunakan adalah data pada tahun 2018 sampai 2022 dengan menggunakan Metode *Smote Extreme Gradient* dengan variabel kanker Serviks, umur, Penggunaan Pembalut, Penggunaan KB Spiral, *Congtraseptive, Pregnacies, smoke* dan Riwayat Keputihan. Hal tersebut menunjukkan bahwa kelas 0 dan 1 sudah bejumlah sama yaitu masing-masing berjumlah 40 data, sehingga dapat dikatakan bahwa masalah ketidakseimbangan data sudah teratasi menggunakan algoritme Smote, sehingga pada hasil tingkat keakuratan AUC metode *Smote extreme gradient boosting* menghasilkan nilai 1.00% yang mana klasifikasi tergolong dalam klasifikasi terbaik.

Pada penelitian ini dapat diketahui bahwa seluruh variabel yang digunakan memperoleh hasil data *Training* dan *Testing* 48 dan 21 dari total data 69, kemudian nilai parameter *Learning_rate* 0.001, 0.01, 0.1 diperoleh nilai parameter *Learning_rate* tebaik yaitu 0.3, dari nilai parameter *Max_depth* 2,3,4,5,6,7diperoleh nilai parameter *Max_depth* terbaik yaitu 6 dan dari nilai parameter *min_child_weight* 1,2,3,4, diperoleh nilai parameter *min_child_weight* terbaik yaitu 1, selanjutnya diperoleh nilai pengujian *SmoteXGBoost: Akurasi 1.00, Presisi 0,97, Recall 0.92, F1-Measure 0.82 dan AUC 1.00.* dimana hasil penelitian tersebut sejalan dengan penelitian [15][5]. Namun hal tersebut berbeda dengan penelitian [16], dimana menunjukkan bahwa terdapat 7 variabel independen dan 1 variabel dependen. Berdasarkan temuan tersebut dapat disimpulkan bahwa penelitian ini telah memperkuat temuan terdahulu dengan hasil penelitian yang relevan. Penelitian ini masih memiliki keterbatasan hanya dilakukan di salah satu RSUD Dr. Pirngadi Kota Medan. Pada penelitian masa mendatang direkomndasikan untuk membandingkan lebih dari satu RSUD serta dengan menggunakan metode yang berbeda lebih dalam lagi.

5. Simpulan

Berdasarkan kesimpulan penelitian yang telah dilakukan bahwa kanker serviks adalah masalah hampir semua ketakutan wanita karena faktor penyebab awalnya sangat sulit sekali dideteksi dengan adanya penerapan metode *smote extreme gradient boosting* ini dapat dilihat bahwa pada faktor gejala pada perhitungan ukur skala menggunakan metode *smote extreme gradient boosting* riwayat keputihan (*Vaginal Discharge*) dengan 82% dan umur (*Age*) diatas 40 tahun dengan 79% lebih besar resikonya terserang penyakit kanker serviks, serta dapat dilihat pada diagram plot jumlah pasien terserang penyakit kanker serviks lebih banyak pada tahun 2018.

Klasikasi pada penyakit kanker serviks dengan menggunakan *SMOTE-XGBoost* memiliki hasil tingkat keakuratan dengan nilai sebesar 1.00% dimana berada dalam rentang 0,91-1.00 sehingga dapat dikatakan bahwa hasil *accuracy* klasifikasi masuk dalam kategori sangat baik dalam mengetahui penyakit kanker serviks

Daftar Referensi

[1] F. Ali, R. Kuelker, and B. Wassie, “Understanding cervical cancer in the context of developing countries,” *Ann. Trop. Med. Public Heal.*, vol. 5, no. 1, pp. 3–15, 2019, doi: 10.4103/1755-6783.92871.

[2] H.P. Samadi, "*Yes, I Know Everything abaut Kanker Serviks! Mengenal, Mencegahnya, Bagaimana Anda Menjalani Pengobatan"*. Solo: Metagraf, PT Tiga Serangkai Pustaka Mandiri, 2018.

[3] Suparyanto, *Kanker Leher Rahim (Carinoma Cervix), 2019*. hhtp://dr.suparyanto. blogspot.com/2011/04/Kanker-Leher-Rahim-Carinoma Cervix. html. (Diakses pada tanggal 9 Januari 2019).

[4] M. A. J. P. American Cancer Society. *Cancer facts & figures 2013. Atlanta: American Cancer Society,* 2019.

[5] N. P. Y. T. Wijayanti, E. N. Kencana, And I. W. Sumarjaya, “Smote: Potensi Dan Kekurangannya Pada Survei,” *E-Jurnal Mat.*, vol. 10, no. 4, p. 235, 2021, doi: 10.24843/mtk.2021.v10.i04.p348.

[6] Kamber, Micheline, Data Mining: Concept and Techniques Second Edition, Morgan Kaufmann Publishers, 2020.

[7] Q. Wang, A hybrid sampling SVM approach to imbalanced data classification. In *Abstract and applied analysis,* Hawai, Vol. 2014.

[8] H.P. Samadi, "*Yes, I Know Everything abaut Kanker Serviks! Mengenal, Mencegahnya,*

Bagaimana Anda Menjalani Pengobatan. Solo: Metagraf, PT Tiga Serangkai Pustaka Mandiri, 2019

[9] J. Brownlee, "Imbalanced Classification with Python". Machine Learning Mastery, 2020.

[10] N. Adhelia, "Implementasi metode *random forest* dan *xgboost* pada klasifikasi *customer churn*. *Skripsi*. Yogyakarta: Fakultas Matematika dan IPA. pp. 38-66, 2020.

[11] T. Chen, & C. Guestrin, "Xgboost: A Scalable Tree Boosting System. In Proceedings Of The 22nd Acm Sigkdd International Conference On Knowledge Discovery And Data Mining, pp. 785-794, 2019.

[12] R. Ubaidillah, M. Muliadi, D. T. Nugrahadi, M. R. Faisal, and R. Herteno, "Implementasi XGBoost Pada Keseimbangan Liver Patient Dataset dengan SMOTE dan Hyperparameter Tuning Bayesian Search," *J. Media Inform. Budidarma*, vol. 6, no. 3, p. 1723, 2022, doi: 10.30865/mib.v6i3.4146.

[13] G. A. Shafila, "Implementasi Metode Extreme Gradient Boosting (Xgboost) untuk Klasifikasi pada Data Bioinformatika (Studi Kasus : Penyakit Ebola , GSE 122692)," *Dspace.Uii.Ac.Id*, pp. 1–77, 2020, [Online]. Available: https://dspace.uii.ac.id/bitstream/handle/123456789/29276/16611022 Gregy Addis Shafila.pdf?sequence=1&isAllowed=y

[14] E. Sri, "Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit". *JOMTA Journal of Mathematics: Theory and Applications* vol. 4, no. 1, pp. 21-26, 2022.

[15] B. Robert and E. B. Brown, "No 主観的健康感を中心とした在宅高齢者における 健康関連指標に関する共分散構造分析Title," no. 1, pp. 1–14, 2022.

[16] S. E. Herni Yulianti, Oni Soesanto, and Yuana Sukmawaty, "Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit," *J. Math. Theory Appl.*, vol. 4, no. 1, pp. 21–26, 2022, doi: 10.31605/jomta.v4i1.1792.

[17] D. M. El Faritsi, D. Saripurna, and I. Mariami, "Sistem Pendukung Keputusan Untuk Menentukan Tenaga Pengajar Menggunakan Metode MOORA," *J. Sist. Inf. Triguna Dharma (JURSI TGD)*, vol. 1, no. 4, p. 239, 2022, doi: 10.53513/jursi.v1i4.4948.

[18] A.N. Rachmi, "Implementasi Metode Random Forest Dan Xgboost Pada Klasifikasi Customer Churn". *Tugas Akhir. Fakultas Matematika Dan Ilmu Pengetahuan Alam Universitas Islam Indonesia Yogyakarta, 2020*.

[19] M. Syukron, R. Santoso, and T. Widiharih, "Perbandingan Metode Smote Random Forest Dan Smote Xgboost Untuk Klasifikasi Tingkat Penyakit Hepatitis C Pada Imbalance Class Data," *J. Gaussian*, vol. 9, no. 3, pp. 227–236, 2020, doi: 10.14710/j.gauss.v9i3.28915.

[20] Yendrizal, "Penentuan Siswa SMK Kimia Analisa Terbaik Yang Akan Dikirim Mengikuti Olimpiade Kimia Tingkat Nasional Menerapkan Metode Entropy dan MOORA," *J. Media Inform. Budidarma*, vol. 4, no. 1, pp. 963–969, 2020, doi: 10.30865/mib.v4i4.2350.

[21] R. Siringoringo, "Klasifikasi data tidak Seimbang menggunakan algoritme SMOTE dan k-nearest neighbor," *J. ISD*, vol. 3, no. 1, pp. 44–49, 2018.

[22] P. Septiana Rizky, R. Haiban Hirzi, and U. Hidayaturrohman, "Perbandingan Metode LightGBM dan XGBoost dalam Menangani Data dengan Kelas Tidak Seimbang," *J Stat. J. Ilm. Teor. dan Apl. Stat.*, vol. 15, no. 2, pp. 228–236, 2022, doi: 10.36456/jstat.vol15.no2.a5548.

[23] M. R. Givari, M. R. Sulaeman, and Y. Umaidah, "Perbandingan Algoritme SVM, Random Forest Dan XGBoost Untuk Penentuan Persetujuan Pengajuan Kredit," *Nuansa Inform.*, vol. 16, no. 1, pp. 141–149, 2022, doi: 10.25134/nuansa.v16i1.5406.

[24] A Damayanti, "LANDASAN TEORI 3.1 Data Mining," pp. 3–13, 2019, [Online]. Available: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwiDsLXgh9X2AhXEQ3wKHcsMAEsQFnoECAIQAQ&url=http%3A%2F%2Feprints.akakom.ac.id%2F6689%2F4%2F4_19780420%2520200501%25202%2520001_BAB_III.pdf&usg=AOvVaw1zmzc5dR6q4_wwT0F6vd

[25] J. Juliono, & D.J. Pasya, "Forcasting Produk Domestik Regional Bruto Atas Dasar Harga Konstan Menurut Pengeluaran Menggunakan Metode Doubel Exponential Smoothing". *Jurnal Ilmiah Ekonomi Manajemen Jurnal Ilmiah Multi Science*, vol. 13, no. 1, pp. 49-57, 2022.

[26] T. T. Maskoen and D. Purnama, "Area Under the Curve dan Akurasi Cystatin C untuk Diagnosis Acute Kidney Injury pada Pasien Politrauma," *Maj. Kedokt. Bandung*, vol. 50, no. 4, pp. 259–264, 2018, doi: 10.15395/mkb.v50n4.1342.