**Journal of Computer System and Informatics (JoSYC)**
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 386−391
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2154

# *Oversampling*, *Undersampling*, Smote SVM dan Random Forest pada Klasifikasi Penerima Bidikmisi Sejawa Timur Tahun 2017

**Laila Qadrini*, Hikmah, Megasari**

Program Studi Statistika, Universitas Sulawesi Barat, Indonesia
Email: laila.qadrini@unsulbar.ac.id
Submitted: **20/08/2022**; Accepted: **30/08/2022**; Published: **30/08/2022**

**Abstrak-**Bidikmisi adalah bantuan biaya pendidikan dari pemerintah bagi lulusan Sekolah Menengah Atas (SMA) atau sederajat yang memiliki potensi akademik baik namun memiliki keterbatasan ekonomi. Berbeda dengan beasiswa yang fokus memberikan penghargaan atau dukungan finansial kepada mereka yang berprestasi. Persyaratan pencapaian Bidikmisi bertujuan agar penerima Bidikmisi dipilih dari mereka yang benar-benar memiliki potensi dan kemauan untuk menyelesaikan pendidikan tinggi. Mengingat penerima bidikmisi ini harus benar-benar orang yang tepat, maka pada penelitian ini akan dilakukan klasifikasi penerima bidikmisi 2017 di Jawa Timur, pada penelitian ini terdapat data yang tidak seimbang kelas "Diterima" yaitu lebih dari kelas "Tidak diterima". Jika data tidak seimbang, hampir semua algoritma klasifikasi akan menghasilkan akurasi yang jauh lebih tinggi untuk kelas mayoritas daripada untuk kelas minoritas. Peneliti akan menangani ketidakseimbangan kelas. Teknik resampling yang digunakan dalam penelitian terkait prediksi penerima bidikmisi meliputi teknik resampling yaitu *Oversampling*, *Undersampling* dan SMOTE dengan menggunakan dua metode klasifikasi yaitu SVM dan *Random Forest*. Teknik *Oversampling* dipilih karena tidak mengurangi jumlah data tetapi menambah dataset yang kurang pada kelas minoritas. Algoritma *Oversampling* yang digunakan adalah *Synthetic Minority Over-sampling Technique* (SMOTE), algoritma ini dipilih dari beberapa algoritma resampling karena SMOTE menghasilkan akurasi yang baik dan efektif dalam menangani kelas yang tidak seimbang karena mengurangi *overfitting*.

**Kata Kunci**: *Oversampling*; *Undersampling*; SMOTE

**Abstract-**Bidikmisi is tuition assistance from the government for high school graduates (SMA) or equivalent who have good academic potential but have economic limitations. Different from scholarships that focus on providing awards or financial support to those who excel. The achievement requirements for Bidikmisi are aimed at ensuring that Bidikmisi recipients are selected from those who truly have the potential and willingness to complete higher education. Given that the recipients of this bidikmisi must really be the right person, in this study a classification of the recipients of the 2017 bidikmisi in East Java will be carried out, in this study there is data that is not balanced the "Accepted" class is more than the "Not accepted" class. If the data is not balanced, almost all classification algorithms will produce much higher accuracy for the majority class than for the minority class. Researchers will handle class imbalances. The resampling technique used in research related to the prediction of bidikmisi recipients includes resampling techniques, namely *Oversampling*, *Undersampling* and SMOTE using two classification methods, namely SVM and Random Forest. The *Oversampling* technique was chosen because it does not reduce the amount of data but adds to the dataset that is lacking in the minority class. The *Oversampling* algorithm used is Synthetic Minority Over-sampling Technique (SMOTE), this algorithm was chosen from several resampling algorithms because SMOTE produces good accuracy and is effective in dealing with unbalanced classes because it reduces overfitting.

**Keywords**: Oversampling; Undersampling; SMOTE

## 1. PENDAHULUAN

Bidikmisi adalah bantuan biaya pendidikan dari pemerintah bagi lulusan Sekolah Menengah Atas (SMA) atau sederajat yang memiliki potensi akademik baik tetapi memiliki keterbatasan ekonomi. Berbeda dari beasiswa yang berfokus pada memberikan penghargaan atau dukungan dana terhadap mereka yang berprestasi [1]. Walaupun demikian, syarat prestasi pada bidikmisi ditujukan untuk menjamin bahwa penerima bidikmisi terseleksi dari yang benar benar mempunyai potensi dan kemauan untuk menyelesaikan pendidikan tinggi. Mengingat bahwa penerima bidikmisi ini mesti benar-benar tepat orang maka pada penelitian ini akan dilakukan klasifikasi penerima bidikmisi Tahun 2017 di Jawa timur, Klasifikasi data mining adalah sebuah proses menemukan definisi kesamaan karakteristik dalam suatu kelompok atau kelas (*class*). Klasifikasi data mining menjadi salah satu metode yang paling umum untuk digunakan. Metode ini dilakukan bertujuan untuk memperkirakan kelas dari suatu objek yang labelnya belum diketahui [2]. SVM Salah satu metode klasifikasi yang cukup terkenal paling kuat dan akurat adalah metode *Support Vector Machine* (SVM) dan *Random Forest*. *Support Vector Machine* (SVM) suatu Teknik untuk melakukan klasifikasi maupun regresi yang sangat popular bekangan ini. SVM berada dalam satu kelas dengan ANN dalam hal fungsi dan kondisi permasalahan yang bisa diselesaikan. *Random forest* bisa lebih efektif mengatasi masalah overfitting, karena ada banyak hasil prediksi yang bisa diambil dan menghilangkan bias yang mungkin ada. Keduanya masuk dalam kelas *supervised learning* [3]. Klasifikasi dapat diterapkan dalam berbagai aspek sehingga seiring berjalannya waktu metode klasifikasi cukup banyak dikembangkan, namun terdapat permasalahan yang sering ditemui dalam klasifikasi yaitu masalah ketidakseimbangan data. Data tidak seimbang merupakan suatu keadaan dimana distribusi kelas data tidak seimbang, jumlah kelas data (*instance*) yang satu lebih sedikit atau lebih banyak dibanding dengan jumlah kelas data lainnya [4]. Kelompok kelas data yang lebih sedikit dikenal dengan kelompok minoritas (*minority*), kelompok kelas data yang lainnya disebut dengan kelompok mayoritas (*majority*). Klasifikasi pada data dengan kelas tidak seimbang merupakan masalah utama pada bidang data mining, misalnya pada masalah medis [5], masalah klasifikasi teks [6], sosial media [7]. Jika bekerja pada data tidak seimbang, hampir semua algoritma klasifikasi akan menghasilkan akurasi yang jauh lebih tinggi untuk

**Journal of Computer System and Informatics (JoSYC)**
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 386−391
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2154

kelas mayoritas daripada kelas minoritas [8]. Perbedaan ini merupakan suatu indikator performa klasifikasi yang buruk. Pada beberapa kasus, kelas minoritas justru lebih penting untuk diidentifikasi daripada kelas mayoritas. Perbandingan antara kelas minoritas dengan kelas mayoritas disebut dengan *Imbalance Ratio* (IR) atau rasio ketidakseimbangan. Semakin besar perbedaan antara kelas minoritas dengan kelas mayoritas maka nilai dari *Imbalance Ratio* (IR) atau rasio ketidakseimbangan semakin besar. Ketidakseimbangan dataset pada data mining adalah masalah yang serius. Dataset yang tidak seimbang menyebabkan misleading atau kesesatan dalam hasil klasifikasi dimana data kelas minoritas sering diklasifikasikan sebagai kelas mayoritas [9]. Penerapan algoritma klasifikasi tanpa memperhatikan keseimbangan kelas mengakibatkan prediksi yang baik bagi kelas mayoritas dan kelas minoritas diabaikan. Apabila algoritma klasifikasi di implementasikan langsung terhadap dataset yang *imbalance* maka akan mengalami penurunan performa [10]. Pada penelitian ini, peneliti akan melakukan penanganan ketidakseimbangan kelas, Teknik resampling yang digunakan pada penelitian terkait prediksi penerima bidikmisi antara lain menggunakan teknik resampling yaitu *Oversampling*, *Undersampling* dan SMOTE menggunakan dua metode klasifikasi yaitu SVM dan *random forest*. Teknik *Oversampling* dipilih karena tidak mengurangi jumlah data akan tetapi menambah dataset yang kurang pada kelas minoritas. Algoritma *Oversampling* yang digunakan adalah *Synthetic Minority Over-sampling Technique* (SMOTE), algoritma ini dipilih dari beberapa algoritma *resampling* karena SMOTE menghasilkan akurasi yang baik dan efektif dalam menangani kelas yang tidak seimbang karena mengurangi overfitting [10]. beberapa penelitian sebelumnya dilakukan oleh [10], Algoritma *Synthetic Minority Over-Sampling Technique* untuk Menangani Ketidakseimbangan Kelas pada Dataset Klasifikasi, Hasil dari penelitian ini dibandingkan dengan hasil klasifikasi tanpa *resampling*. Uji evaluasi yang digunakan ialah akurasi, *Geometric Mean*(g-mean), dan *Confussion Matrix* (CM). Penanganan distribusi kelas yang tidak seimbang pada dataset menggunakan algoritma SMOTE dapat meningkatkan nilai akurasi maupun g-mean pada algoritma Naïve Bayes, SVM, KNN dan *Decision Tree*. Beberapa penelitian tentang klasifikasi *imbalanced* data yang telah dilakukan dengan menggunakan metode SMOTE adalah [11], melakukan klasifikasi kerentanan seseorang terserang penyakit stroke di Jawa Timur dengan menggunakan SMOTE dan SVM (*Support Vector Machine*). Penelitian tersebut menggunakan data sebanyak 65918 observasi, rasio penderita stroke dan bukan penderita stroke adalah 1:129. Setelah dilakukan SMOTE, proporsi data menjadi seimbang dan didapatkan hasil ketepatan klasifikasi akurasi, sensitivitas, dan spesifitas yang lebih tinggi daripada data stroke awal yang *imbalanced*. penelitian yang dilakukan [12] adalah menganalisis efektifitas SMOTE dalam meningkatkan ketepatan akurasi klasifikasi. Data yang digunakan adalah desa 5 Kabupaten di Jawa Timur yang berjumlah 1.122 desa dengan kelompok berstatus desa tertinggal sebanyak 115 desa. hasilnya Peningkatan nilai G-*mean* dan sensitivitas tertinggi pada kombinasi SMOTE dan LR Ridge dengan semua variabel. adapun penelitian yang dilakukan [13] menghasilkan model klasifikasi yang baik untuk melakukan prediksi kebangkrutan. *Resampling* diterapkan pada data latih agar menghasilkan model klasifikasi yang lebih optimal. Metode *resampling* yang digunakan adalah kombinasi SMOTE dan *Undersampling*. Metode klasifikasi yang digunakan untuk prediksi adalah *multilayer perceptron* dan *complement naïve bayes*. Performa prediksi dihitung menggunakan skor *recall*, ROC AUC, dan PR AUC. Berdasarkan hasil pengujian, penggunaan SMOTE dan *Undersampling* cukup signifikan dalam memperbaiki model klasifikasi pada *multilayer perceptron*.

# 2. METODOLOGI PENELITIAN

## 2.1 Data Penelitian

Dataset yang digunakan di penelitian ini adalah data penerima Bidikmisi seJawa timur Tahun 2017, Total data pada penelitian ini adalah 52420.

**Tabel 1**. Deskripsi Variabel Penelitian

| Variabel | Nama Variabel | Keterangan |
|---|---|---|
| $Y$ | Status Penerimaan Beasiswa Bidikmisi | 0= Tidak Diterima Bidikmisi<br>1=Diterima Bidikmisi |
| $X_1$ | Pekerjaan Ayah | 1=Tidak Bekerja<br>2=Petani, Nelayan, Lainnya<br>3=TNI/POLRI<br>4=Wirausaha<br>5=Peg.Swasta<br>6=PNS |
| $X_2$ | Pekerjaan Ibu | 1=Tidak Bekerja<br>2=Petani, Nelayan, Lainnya<br>3=TNI/POLRI<br>4=Wirausaha<br>5=Peg.Swasta |

**Journal of Computer System and Informatics (JoSYC)**
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 386−391
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2154

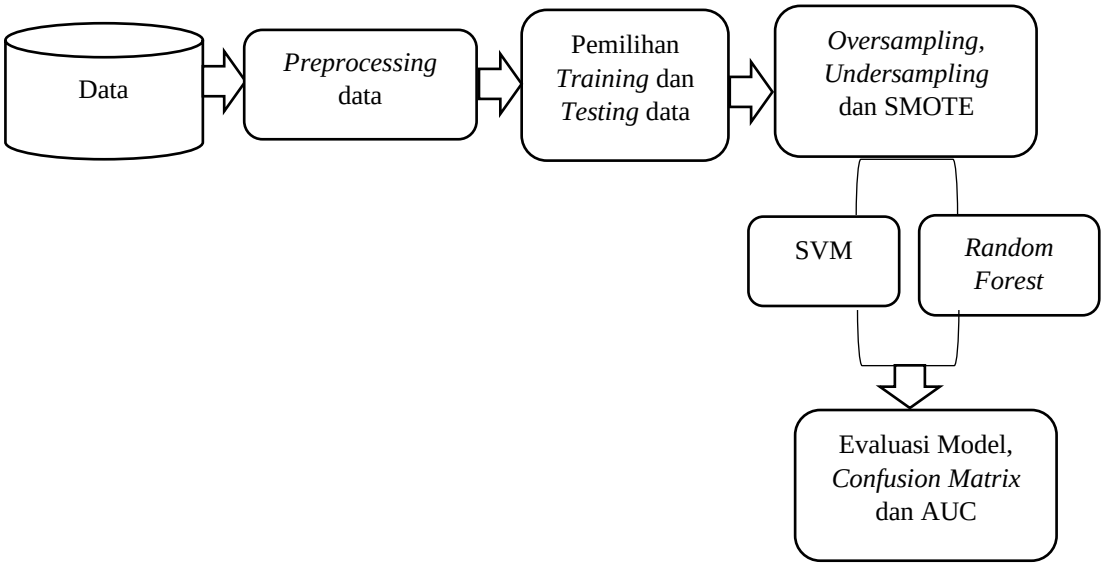
| Variabel | Nama Variabel | Keterangan |
|---|---|---|
| | | 6=PNS |
| $X_3$ | Pendidikan Ayah | 1=Tidak Sekolah<br>2=Pendidikan Dasar (SD/MI dan SMP/MTs)<br>3=SMA/MA<br>4=PT (D1,D2,D3,D4/S1)<br>5=PT (S2,S3) |
| $X_4$ | Pendidikan Ibu | 1=Tidak Sekolah<br>2=Pendidikan Dasar (SD/MI dan SMP/MTs)<br>3=SMA/MA<br>4=PT (D1,D2,D3,D4/S1)<br>5=PT (S2,S3) |
| $X_5$ | Penghasilan Ayah | 1=Tidak Berpenghasilan<br>2=< Rp.1000.000<br>3=Rp.1000.000 - Rp.2000.000<br>4=Rp.2000.001- Rp.300.000<br>5= > Rp. 3000.001 |
| $X_6$ | Penghasilan Ibu | 1=Tidak Berpenghasilan<br>2=< Rp.1000.000<br>3=Rp.1000.000 - Rp.2000.000<br>4=Rp.2000.001- Rp.300.000<br>5= > Rp. 3000.001 |
| $X_7$ | Kepemilikan Rumah Tinggal Keluarga | 1=Tidak Memiliki Rumah<br>2=Sewa(Tahunan, Bulanan), menumpang, dan menumpang tanpa ijin<br>3=Sendiri |
| $X_8$ | Sumber Listrik yang digunakan Keluarga | 1=Tidak Ada<br>2=Genset/Mandiri, Tenaga Surya<br>3=PLN |
| $X_9$ | Luas Tanah Rumah Tinggal Keluarga | 1=25-50 m<br>2=50-9m<br>3=>100 m |
| $X_{10}$ | Luas Bangunan Rumah Tinggal Keluarga | 1=25-50 m<br>2=50-99m<br>3=>100 m |
| $X_{11}$ | Kepemilikan Fasilitas Mandi Cuci Kakus | 1=Berbagi Pakai<br>2=Kepemilikan sendiri didalam<br>3=Kepemilikan sendiri diluar |
| $X_{12}$ | Jumlah Tanggungan | 1 =< 1<br>2=2<br>3=3<br>4=>4 |

### 2.2 Kerangka Kerja Penelitian

Penelitian ini menerapkan dua metode klasifikasi yaitu metode SVM dan Random Forest. Pada Seleksi Fitur digunakan untuk menyeleksi data yang rusak/tidak lengkap menggunakan fitur "*Input Missing Value*" dan "*Rename Unused Value*" dengan menggunakan metode SVM sehingga didapatkan data set murni. Mentranformasikan data dari numerik ke nominal dan lakukan normalisasi menentukan bentuk data yang paling tepat. Selanjutnya membagi data *training* dan data *testing* dengan perbandingan 75%:25%. Pemrosesan data dilakukan dengan melakukan metode SVM dan *Random Forest* serta penerapan *Oversampling*, *Undersampling* dan SMOTE pada pengkalsfikasinya, dalam proses validasi sehingga diperoleh akurasi yang lebih tinggi dari pengolahan masing-masing metode. Penelitian ini dilakukan untuk pengembangan evaluasi dan pemecahan

**Journal of Computer System and Informatics (JoSYC)**
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 386−391
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2154

masalah yaitu meningkatkan akurasi pada pelaksanaan program Bidikmisi agar tepat sasaran. Berikut ini adalah tahapan penelitian yang dilakukan:

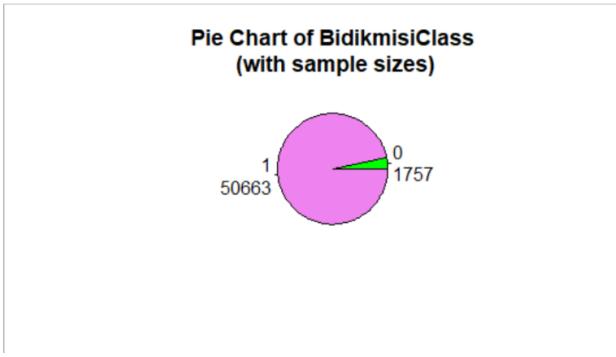



**Gambar 1**. Tahapan Penelitian

## 3. HASIL DAN PEMBAHASAN

### 3.1 Gambaran Umum Data

Data memiliki 13 atribut dimana salah satu atributnya adalah label kelas. Kelas tidak diterima sebagai kelas minoritas memiliki 1757 data dan kelas diterima sebagai kelas mayoritas memiliki 50663 data, dengan proporsional kelas minoritas dan mayoritas adalah 0,03:0,97.




**Gambar 2**. Sebaran Kelas “Ya” dan “Tidak”

### 3.2 *Preprocessing* data

*Preprocessing* bertujuan untuk menyeragamkan rentang data antara 0 sampai 1. Agar nilai atribut antara 0 dan 1, maka setiap nilai atribut dibagi dengan nilai maksimal atribut dari semua data. Untuk atribut bidikmisi*class*, yang sebelumnya tipe *string* maka dilakukan *recode* dan terdeteksi tipe integer, selanjutnya mengubah tipe numerik menjadi tipe kategorik yaitu 0 untuk tidak dan 1 untuk Ya, sedangkan untuk atribut yang lain tetap bernilai *integer*.

### 3.3 Seleksi Fitur

Seleksi fitur adalah salah satu aktivitas yang paling penting dilakukan dalam membangun model dan sering dilakukan berulang-ulang. Tidak semua fitur pada penelitian ini bisa bermanfaat dalam pembuatan model, bahkan bila terlalu banyak fitur yang tidak relevan pada data dapat menurunkan akurasi model. Kita harus memilih fitur yang meberikan hasil yang terbaik, pada penelitian ini seleksi fitur telah dilakukan dan tereduksi menjadi 13 fitur.

### 3.3 Pemilihan *training* dan *testing* data

Agar dapat melakukan prediksi, data harus dibagi terlebih dahulu ke dalam data *training* dan data *testing*. Pada penelitian ini keseluruhan data dibagi menjadi 75%:25% untuk 75% data *training* dan 25% data *testing*. Tabel 2 menunjukkan jumlah data *training* dan data *testing*.

**Journal of Computer System and Informatics (JoSYC)**
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 386−391
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2154

**Tabel 2**. Pembagian data *training* dan data *testing*

| Jenis Data | Jumlah data | | Total |
|---|---|---|---|
| | Kelas 1 | Kelas 0 | |
| *Training* | 37988 | 1327 | 39315 |
| *Testing* | 12675 | 430 | 13105 |

### 3.4 Hasil Prediksi SVM tanpa *Handling*, *Oversampling*, *Undersampling*, dan SMOTE SVM

Terdapat beberapa pendekatan untuk mengatasi masalah *imbalanced* data, yaitu pendekatan pada level data dengan teknik pengambilan sampel, pendekatan level algoritma, serta metode ensemble [14]. Teknik pengambilan sampel yang biasanya digunakan untuk mengatasi masalah *imbalanced* data yaitu *over-sampling*, *under-sampling*, dan kombinasi keduanya. Hasil dari *random Oversampling* tidak selalu meningkatkan prediksi kelas minor. Apabila data kelas minor diduplikasi dalam jumlah yang besar, maka akan sulit untuk mengidentifikasi data yang memiliki kemiripan karakteristik namun berada di kelas yang berbeda. Berkebalikan dengan *Oversampling*, *Undersampling* adalah metode untuk mengambil beberapa data mayoritas sehingga jumlah data mayoritas sama besar jumlahnya dengan jumlah data minoritas [15]. *Synthetic Minority Oversampling Technique* (SMOTE) memilih data pada kelas minoritas secara acak kemudian mencari k data kelas minoritas terdekat dari data tersebut [16]. Dari k data kelas minoritas tersebut, akan dipilih satu data secara acak yang selanjutnya dihubungkan dengan data awal yang dipilih untuk membentuk segmen garis pada ruang fitur. Data sintesis dihasilkan menggunakan kombinasi *convex* antara dua data yang dipilih secara acak tadi. Pendekatan ini dinilai cukup efektif karena data sintesis yang dibentuk relatif dekat dalam ruang fitur dengan data yang ada dari kelas minoritas. Prosedur ini dapat digunakan untuk membuat sebanyak mungkin data sintetik untuk kelas minoritas yang diperlukan.

**Tabel 3**. Hasil Prediksi SVM, *Oversampling*, *Undersampling*, dan SMOTE

| Pengukuran | Data *Training* tanpa *Handling* | Data *Training* setelah *Oversampling* | Data *Training* setelah *Undersampling* | Data *Training* setelah SMOTE |
|---|---|---|---|---|
| Akurasi | 0.9672 | 0,591 | 0,032 | 0,761 |
| Presisi | 1 | 0,488 | 1 | 0,237 |
| *Recall* | 0 | 0,595 | 0 | 0,779 |
| AUC | 0.5 | 0,542 | 0,5 | 0,508 |

Akurasi model diatas bernilai 96,72% yang berarti sudah sangat baik, secara umum dibidang data sains model dengan akurasi diatas 70% dapat digolongkan sebagai model yang berkinerja cukup baik [18]. Namun model SVM tanpa *handling imbalanced* ini memiliki sensitivitas yang rendah yaitu 0, 00%. Karena tidak dapat mendeteksi satupun kelas "Ya". Akurasi ini mengukur keseluruhan akurasi model, tanpa membedakan *error* FP ataupun FN. Informasi akurasi ini sebenarnya kurang informatif terutama pada penerapan model yang lebih difokuskan pada mendeteksi hal-hal yang sangat peka pada *false positive* atau *false negative* saja. *Precision* memberi petunjuk seberapa baik model dapat memprediksi yang positif. nilai presisi klasifikasi *random sampling* dengan *Oversampling* sebesar 0,488 hal ini menunjukkan hanya 48% model berhasil memprediksi data yang positif, nilai presisi setelah *Undersampling* sebesar 100%, serta nilai presisi setelah SMOTE juga sebesar 23,7%. Sensitifitas atau *recall* mengukur banyaknya data yang yang sukses diprediksi sebagai positif dibandingkan dengan seluruh data yang pada kenyataannya positif. Model menunjukkan angka *sensitivity* untuk *Oversampling* sebesar adalah 59,5%, *Undersampling* sebesar 0% dan SMOTE sebesar 77,9%. Adapun nilai AUC dipakai sebagai ukuran baik buruknya suatu model [19]. AUC mendekati 1 menunjukkan model yang mendekati sempurna, sementara AUC mendekati 0,5 adalah model yang buruk [18]. Nilai AUC untuk model *Oversampling* adalah 0,542, model *Undersampling* 0,5 dan model SMOTE SVM 0,508.

### 3.5 Hasil Prediksi dengan *Random Forest* tanpa *Handling, Oversampling, Undersampling*, dan SMOTE

**Tabel 4**. Hasil Prediksi *Random Forest*, *Oversampling*, *Undersampling*, dan SMOTE

| Pengukuran | Data *Training* tanpa *Handling* | Data *Training* setelah *Oversampling* | Data *Training* setelah *Undersampling* | Data *Training* setelah SMOTE |
|---|---|---|---|---|
| Akurasi | 0,9672 | 0,735 | 0,199 | 0,78 |
| Presisi | 1 | 0,323 | 0,832 | 0,258 |
| *Recall* | 0 | 0,749 | 0,177 | 0,798 |
| AUC | 0,5 | 0,536 | 0,504 | 0,528 |

Akurasi model diatas bernilai 96,7% yang berarti sudah sangat baik seperti model pada SVM, namun model Random forest tanpa *handling imbalanced* juga kurang baik karena nilai sensitivitasnya 0,00%. Artinya model tidak satupun dapat mendeteksi kelas "Ya". Hal ini juga kurang informatif untuk memprediksi *false positive* atau *false negative* saja. Nilai presisi klasifikasi *random sampling* dengan *Oversampling* sebesar 0,323 hal ini menunjukkan hanya 32,3% model berhasil memprediksi data yang positif, nilai presisi setelah *Undersampling* sebesar 0,832 atau 83,2%, serta nilai presisi setelah SMOTE juga sebesar 0,258 atau 25,8%. Model menunjukkan

**Journal of Computer System and Informatics (JoSYC)**
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 386−391
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2154

angka *sensitivity* untuk *Oversampling* sebesar adalah 0,749 atau 74,9%, *Undersampling* sebesar 0,177 atau 17,7% dan SMOTE sebesar 0,798 atau 79,8. Nilai AUC untuk model *Oversampling random forest* adalah 0,536, model *Undersampling random forest* 0,504 dan model SMOTE sebesar 0,508. Nilai presisi yang tinggi belum tentu nilai sensitivitas juga tinggi, demikian pula sebaiknya. Pada penelitian ini nilai AUC tidak mendekati 1 untuk semua model, jika nilai AUC mendekati 1 bisa jadi model menjadi *overfit* yaitu hanya bagus ketika diukur dengan *training* data.

## 4. KESIMPULAN

Hasil prediksi dengan *classifier* SVM dan *Random Forest* tanpa *Handling* memberikan akurasi yang tinggi sebesar 96,72%, namun sebenarnya nilai sensitivitasnya 0 artinya model kurang baik karena tidak mampu memprediksi kelas "Ya". Evaluasi model pada *classifier* SVM dan *Random Forest* setelah *Oversampling* dan SMOTE hampir sama. Pada penelitian ini dapat disimpulkan bahwa penerapan *random sampling Oversampling* dan SMOTE memberikan nilai AUC yang hampir sama dan dapat diterapkan untuk kasus data tak seimbang karena menyebabkan nilai akurasi, presisi, recall dan AUC yang tinggi, tidak *overfit* ataupun *underfit*.

## REFERENCES

[1] https://bidikmisi.belmawa.ristekdikti.go.id/ diakses pada Tanggal 11 Juni 2022.
[2] Tan, P., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining*. Boston: Pearson Education.
[3] Santosa, B. (2007). *Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Yogyakarta(ID): Graha Ilmu.
[4] Ali, S. M. Shamsuddin, & A. L. Ralescu. (2009). Classification with class imbalance problem: a review. *Int J Adv. Soft Compu Appl*, 7(3).
[5] Kothan. (2015). Handling class imbalance problem in miRNA dataset associated with cancer. *Bioinformation*, 11(1):6–10.
[6] Wu, Y. Ye, H. Zhang, M. K. Ng, & S.-S. Ho. (2014). ForesTexter: An efficient random forest algorithm for imbalanced text categorization. *Knowl-Based Syst*. 67:105–116.
[7] Li & S. Liu. (2014). A comparative study of the class imbalance problem in Twitter spam Detection. *Concurr. Comput. Pract*. Exp.,pp. n/a-n/a
[8] Siringoringo, Rimbun. (2018). Klasifikasi data tidak seimbang menggunakan algoritma smote dan k-nearest neighbor. *Jurnal ISD*. 3(1): 2528-5114.
A. Smote & D. A. N. Neighbor. (2017). Klasifikasi Data Tidak Seimbang.3(1):44–49.
[9] M. Mustaqim, B. Warsito, & B. Surarso. (2019). Kombinasi Synthetic Minority *Oversampling* Technique (SMOTE) dan Neural Network Backpropagation untuk menangani data tidak seimbang pada prediksi pemakaian alat kontrasepsi implan. *Regist. J. Ilm. Teknol. Sist. Inf*. 5(2):128.
[10] Sulistiyono, M., Pristyanto, Y., Adi, S., Gumelar, G. (2021). Implementasi Algoritma Synthetic Minority Over-Sampling Technique untuk Menangani Ketidakseimbangan Kelas pada Dataset Klasifikasi. *SISTEMASI: Jurnal Sistem Informasi*. 10(2):445-459.
[11] Novritasari, A. A., & Purnami, S. W. (2015). Klasifikasi Kerentanan Seseorang Terserang Stroke di Jawa Timur Menggunakan Synthetic Minoryty *Oversampling* Technique (SMOTE) dan Support Vector Machine (SVM). Surabaya: *Tugas Akhir*. ITS.
[12] Imanwardhani, C.S. (2018). Pendekatan synthetic minority *Oversampling* technique dalam menangani klasifikasi imbalanced data biner (studi kasus: status ketertinggalan desa di Jawa timur). Surabaya: *Tugas Akhir*. ITS.
[13] Sabilla, I. W., Vista, B. C., (2021). Implementasi SMOTE dan Under Sampling pada Imbalanced Dataset untuk Prediksi Kebangkrutan Perusahaan. *Jurnal Politeknik Caltex Riau*. 7(2):329-339.
[14] Choi, M. J. (2010). A Selective Sampling Method for Imbalanced Data Learning on Support Vector Machines. Iowa: *Graduate Theses*. Iowa State University.
[15] Yen, S.-J., & Lee, Y.-S. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, 36(3): 5718–5727.
[16] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal Of Artificial Intelligence Research*. 16:321-357.
[17] Kurniawan, Dios. (2020). *Pengenalan Machine Learning dengan Python Solusi Untuk Permasalahan Bigdata*. Jakarta(ID): PT. Elex Media Komputindo.
[18] Qadrini, L. Seppewali, A, Aina, A. (2021). Decision tree dan adaboost pada klasifikasi penerima program bantuan sosial. *Jurnal Inovasi Penelitian*. 2(7): 2722-9475.