**MENGKOMBINASIKAN TEKNIK RESAMPLING DENGAN ALGORITMA MACHINE LEARNING UNTUK MENGATASI DATASET TAK SEIMBANG**

**SKRIPSI**



***Oleh :***

**ZINEDINE KAHLIL GIBRAN ZIDANE**

**H13116304**

**PROGRAM STUDI ILMU KOMPUTER**

**DEPARTEMEN MATEMATIKA**

**FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM**

**UNIVERSITAS HASANUDDIN**

**2019**

**MENGKOMBINASIKAN TEKNIK RESAMPLING DAN ALGORITMA MACHINE LEARNING UNTUK MENGATASI DATASET TAK SEIMBANG**

**SKRIPSI**



Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer pada Program Studi Ilmu Komputer Departemen Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin Makassar

**ZINEDINE KAHLIL GIBRAN ZIDANE**

**H13116304**

**PROGRAM STUDI ILMU KOMPUTER**

**DEPARTEMEN MATEMATIKA**

**FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM**

**UNIVERSITAS HASANUDDIN**

**2019**

(LEMBAR PERNYATAAN KEOTENTIKAN)

(DISETUJUI OLEH)

HALAMAN PENGESAHAN

KATA PENGANTAR

(PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIK)

ABSTRAK

Sebagian besar algoritma klasifikasi pada domain data sains yang ada mengasumsikan data training terdistribusi secara seimbang, sementara sebagian besar *real-life* dataset terdistribusi secara tak seimbang. Pada kasus kelas yang tidak seimbang, jumlah *record* suatu kelas (mayoritas) jauh melampaui kelas yang lain (minoritas), namun kelas minoritas sering lebih menarik dan lebih penting untuk diidentifikasi. Dalam masalah distribusi kelas yang tidak seimbang ini, klasifikasi yang benar untuk sampel kelas minoritas lebih berharga daripada kelas mayoritas. Namun, karena distribusi data tidak seimbang, algoritma klasifikasi memiliki kesulitan untuk mengklasifikasi sampel kelas minoritas dengan benar. Jika kinerja model algoritma baik untuk sampel kelas mayoritas namun buruk untuk kelas minoritas maka masalah ketidakseimbangan ini adalah suatu hal yang krusial untuk ditangani. Banyak solusi yang ditawarkan untuk masalah ini, yaitu dengan teknik oversampling pada kelas minoritas dan/atau teknik undersampling pada kelas mayoritas. Pada penelitian ini, penulis mencoba berbagai macam teknik sampling dan mengujinya pada berbagai algoritma klasifikasi machine learning untuk mengetahui kombinasi teknik sampling dan algoritma yang memiliki akurasi tertinggi dalam mengklasifikasi sampel kelas minoritas dan tetap mempertimbangkan klasifikasi kelas mayoritas.

FIXME

ABSTRACT

Table of Contents

[(LEMBAR PERNYATAAN KEOTENTIKAN) 3](#_Toc17530174)

[(DISETUJUI OLEH) 4](#_Toc17530175)

[HALAMAN PENGESAHAN 5](#_Toc17530176)

[KATA PENGANTAR 6](#_Toc17530177)

[(PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIK) 7](#_Toc17530178)

[ABSTRAK 8](#_Toc17530179)

[ABSTRACT 9](#_Toc17530180)

[Table of Contents 10](#_Toc17530181)

[Gambar 13](#_Toc17530182)

[Persamaan 14](#_Toc17530183)

[Tabel 15](#_Toc17530184)

[Lampiran 16](#_Toc17530185)

[BAB I PENDAHULUAN 17](#_Toc17530186)

[1.1 Latar Belakang 17](#_Toc17530187)

[1.2 Rumusan Masalah 21](#_Toc17530188)

[1.3 Batasan Masalah 21](#_Toc17530189)

[1.4 Tujuan Penelitian 22](#_Toc17530190)

[1.5 Manfaat Penelitian 22](#_Toc17530191)

[1.6 Organisasi Skripsi 22](#_Toc17530192)

[BAB II TINJAUAN PUSTAKA 23](#_Toc17530193)

[2.1 Landasan Teori 24](#_Toc17530194)

[2.1.1 Teknik Resampling 24](#_Toc17530195)

[2.1.2 Algoritma Klasifikasi 24](#_Toc17530196)

[2.2 Penelitian Terkait 24](#_Toc17530197)

[2.2.1 Influence of Resampling on Accuracy of Imbalanced Classification 24](#_Toc17530198)

[2.2.2 Comparing oversampling techniques to handle the class imabalance problem: A customer churn prediction case study 24](#_Toc17530199)

[2.2.3 An approach for classification of highly imbalanced data using weighting and undersampling 24](#_Toc17530200)

[2.2.4 A study of the Behavior of Several Methods for Balancing Machine Learning Training Data 24](#_Toc17530201)

[BAB III METODE PENELITIAN 26](#_Toc17530202)

[2.3 Waktu dan Tempat 26](#_Toc17530203)

[2.4 Tahapan Penelitian 26](#_Toc17530204)

[2.5 Sumber Data 26](#_Toc17530205)

[2.6 Instrumen Penelitian 26](#_Toc17530206)

[BAB III HASIL DAN PEMBAHASAN 27](#_Toc17530207)

[3.1 asd 27](#_Toc17530208)

[3.2 ssd 27](#_Toc17530209)

[3.3 asdf 27](#_Toc17530210)

[BAB IV KESIMPULAN DAN SARAN 28](#_Toc17530211)

[4.1 Kesimpulan 28](#_Toc17530212)

[4.2 Saran 28](#_Toc17530213)

[Daftar Pustaka 29](#_Toc17530214)

[LAMPIRAN 37](#_Toc17530215)

Gambar

No table of figures entries found.

Persamaan

**No table of figures entries found.**

**sdasdasd**

Tabel

No table of figures entries found.

Lampiran

**No table of figures entries found.**

# PENDAHULUAN

## Latar Belakang

Dalam beberapa tahun terakhir, perkembangan yang pesat dalam sains dan teknologi telah berdampak pada pertumbuhan data mentah secara eksponensial [1]. Berdasarkan World Economic Forum [2], data digital di dunia akan mencapai 44 zettabyte, atau 44 triliun gigabyte pada tahun 2020. Jumlah ini akan terus berkembang hingga lebih dari dua kali lipat setiap dua tahunnya [3, 4].

Dari pertumbuhan data tersebut, kebutuhan untuk menganalisis data terus meningkat [5]. Menurut laporan dari LinkedIn pada tahun 2018, permintaan pekerjaan yang membutuhkan analisis data berkembang hingga 12 kali lipat lebih banyak dari tahun 2014 [6]. Ini disebabkan karena data merupakan kunci dari setiap organisasi, institusi atau perusahaan untuk mengambil keputusan yang lebih cermat dan efektif [7].

Namun data mentah atau *raw data* tidak dapat memberikan informasi yang berguna. Data tersebut perlu diolah dan dianalisis menjadi informasi yang berguna. *Raw data* yang dimaksud adalah data yang belum diproses, yang terstruktur dan yang tidak terstruktur. Data yang terstruktur adalah data yang disimpan dengan format yang telah ditentukan seperti database [8], di mana atribut data dapat dibedakan dengan jelas sehingga dapat langsung diproses oleh peralatan komputasi [9]. Sedangkan data tidak terstruktur adalah data yang disimpan dalam format yang tidak terstruktur, sehingga membutuhkan campur tangan manusia agar dapat diinterpretasi oleh mesin; seperti dokumen, gambar, video dan audio [10]. Menurut laporan Beal [8], 80% hingga 90% data di dunia tidak terstruktur.

Dengan jumlah data yang sangat banyak, tidak mungkin oleh manusia untuk menganalisis dan membuat perhitungan mengenai data secara manual. Maka dari itu diperlukan bidang khusus untuk mengolah dan menganalisis data. *Data Science* adalah bidang yang mempelajari bagaimana meng-ekstraksi *raw data* menjadi *meaningful information* atau informasi yang berguna [11, 12]. Data Science melibatkan prinsip, proses, dan teknik untuk memahami fenomena melalui data [13]. Data Science merupakan bidang yang sangat luas dan sedang dikembangkan [13], namun salah satu bidang khusus dari data science adalah *machine learning* yang merupakan perpotongan antara *computer science* (ilmu komputer) dan *statistics* (statistika) [14]. Machine learning membahas mengenai bagaimana membangun sistem komputer yang dapat belajar melalui pengalaman tanpa harus diprogram secara spesifik dan manual [14] [15].

Untuk menyelesaikan suatu masalah pada komputer, dibutuhkan algoritma. Algoritma adalah kumpulan instruksi-instruksi yang berurutan yang digunakan untuk membawa suatu input menjadi output tertentu [16]. Namun untuk beberapa masalah, kita tidak mempunyai algoritmanya. Contohnya adalah bagaimana komputer mengenali angka dalam bentuk tulisan tangan [17] dan mengklasifikasi suatu email menjadi spam atau bukan spam [16]. Dalam hal ini, kita dapat mengumpulkan seluruh email yang berlabel spam dan yang bukan spam dan “mempelajari” apa yang membedakan mereka. Di dalam hal ini lah machine learning bekerja.

Data yang dikumpulkan biasanya dalam bentuk dataset atau tabel, di mana setiap kolomnya adalah atribut atau ciri-ciri dan setiap barisnya adalah instansi atau observasi. Dataset tersebut ada yang memiliki kolom label, atau kolom yang berisi informasi mengenai kategori dari setiap observasi (contohnya, spam atau bukan spam), dan ada juga dataset yang tidak memiliki kolom label, di mana isinya hanyalah atribut atau ciri-ciri dari setiap observasi, tanpa mengindikasikan kategori dari tiap observasi [17]. Pembelejaran pada dataset berlabel disebut *supervised learning*. Kasus di mana tujuannya adalah mengklasifkasikan input data ke suatu kategori diskrit tertentu disebut *klasifikasi*, dan kasus di mana outputnya adalah suatu variabel kontinu disebut *regresi*. Selain itu, pembelajaran pada dataset tanpa label atau acuan kategori yang benar disebut *unsupervised learning*. Kasus *unsupervised learning* di mana tujuannya adalah mengelompokkan observasi-observasi yang mirip disebut *clustering*, jika menentukan distribusi data pada input disebut *estimasi kepadatan*. Dan yang terakhir, pembelajan di mana mesin dilatih untuk membuat keputusan tertentu dengan cara *trial and error* diseut *reinforcement learning* [18]. Masing-masing jenis pembelajaran memiliki banyak algoritma yang telah dikembangkan dengan berbagai pendekatan yang berbeda-beda [14]. Berdasarkan dokumentasi dari *scikit-learn* [19], terdapat lebih dari 100 algoritma machine learning yang ada.

Untuk distribusi data pada suatu dataset, terdapat istilah kelas yang terdistribusi secara seimbang (*balanced*) dan secara tak seimbang (*imbalanced*). Dataset dengan kelas yang seimbang berarti jumlah observasi untuk setiap kelas tidak jauh dari kelas-kelas yang lain [20]. Sedangkan untuk dataset dengan distribusi kelas yang tak seimbang, jumlah suatu observasi pada kelas tertentu sangat jauh berbeda dengan kelas yang lain. Hal ini berlaku pada dataset dengan kelas biner (dua kelas saja) dan juga *multiclass* (lebih dari dua kelas) [1]. Kelas dengan jumlah observasi sedikit disebut kelas minoritas (*minority* *class*) dan kelas dengan jumlah observasi yang sangat banyak disebut kelas mayoritas (*majority* *class*). Tidak jarang suatu dataset terdistribusi secara tak seimbang dengan proporsi antara kelas minoritas dan mayoritasnya adalah 1:100, 1:1000, atau 1:1000 [1]. Sebagian besar data asli di dunia terdistribusi secara tak seimbang [1, 21, 22, 23].

Pada umumnya, algoritma-algoritma machine learning, dalam hal ini pada masalah klasifikasi, bekerja dengan tujuan utama memaksimalkan akurasi [24]. Hal ini sangat masuk akal, karena akurasi yang tinggi menjelaskan bahwa model algoritma tersebut melaksanakan tugasnya dengan baik, mengklasifikasikan kategori data dengan benar dengan sedikit kesalahan. Namun, akurasi hanya memberikan informasi secara umum, bagaimana jika model algoritma tersebut bekerja pada dataset tak seimbang, dan hanya mampu mengklasifikasikan kelas mayoritas dengan benar tetapi tak mampu mengklasifikasikan kelas minoritas? Jika perbandingan antara kelas minoritas dan mayoritas saja satu berbanding seratus, maka kita akan mendapatkan akurasi lebih besar dari 99%, dengan kesalahan lebih kecil dari 1% yang hampir seluruhnya adalah kelas minoritas. Masalah ini memberi bias terhadap performa algoritma-algoritma klasifikasi, terutama jika kelas yang lebih utama untuk diklasifikasikan dengan benar adalah kelas minoritas, seperti email spam, diagnosis penyakit di bidang kedokteran, deteksi kartu kredit palsu dan lain-lain [23, 25]. Hal ini menunjukkan bahwa pada kasus dataset tak seimbang, dibutuhkan perhatian lebih terhadap *preprocessing* data sebelum dimasukkan ke model.

Banyak cara yang telah ditemukan untuk mengatasi dataset tak seimbang ini, seperti melakukan *resampling* terhadap data yang ada. Resampling adalah teknik mengambil sampel secara berulang dari sampel data asli [26]. Teknik resampling terdiri dari *oversampling*, yaitu mengambil sampel berulang kali dari kelas minoritas; dan undersampling, yaitu mengambil sampel secara acak dari kelas mayoritas [27]. Kedua teknik ini dapat digunakan secara terpisah ataupun digabung [27, 28, 29, 30]. Masing-masing teknik resampling memiliki metode yang berbeda-beda, dengan performa yang berbeda pula [27, 29]. Begitu juga dengan algoritma klasifikasi pada machine learning, terdapat banyak metode yang berbeda dengan performa yang berbeda pula [31].

Dalam beberapa penelitian terkait [29, 32, 33, 27], telah dilakukan berbagai percobaan untuk mengatasi masalah dataset tak seimbang, namun metode-metode resampling maupun algoritma machine learning yang digunakan tidak beragam untuk mengetahui metode terbaik untuk mengatasi masalah ini. Seperti penelitian yang dilakukan Amin dkk. [33] hanya meneliti teknik oversampling, [27, 29, 32] meneliti teknik oversampling dan undersampling namun hanya menggunakan satu algoritma machine learning, sedangkan Diri [31] hanya meneliti beberapa algoritma machine learning tanpa pertimbangan dataset tak seimbang. Sedangkan untuk mengetahui metode resampling dan algoritma machine learning terbaikuntuk masalah ini, dibutuhkan kombinasi-kombinasi antar teknik resampling, dan juga antar algoritma machine learning. Setiap kombinasi (pasangan) ini, seperti SMOTE dengan Support Vector Machine, atau Tomek Links dengan Naive Bayes Classifier akan diuji performanya terhadap dataset yang diberikan, kemudian dari kombinasi-kombinasi tersebut dapat ditarik kesimpulan mengenai kombinasi algoritma dan teknik resampling yang terbaik, dan algoritma machine learning dengan performa terbaik, dan teknik resampling dengan performa terbaik. Setiap kombinasi atau pasangan dievaluasi hasilnya dengan tidak hanya pada satu dataset tak seimbang saja, melainkan dengan beberapa dataset tambahan untuk mendapatkan hasil yang lebih umum dan tanpa bias.

Berdasarkan uraian di atas, penulis ingin melakukan penelitian mengenai dataset tak seimbang dengan cara “Mengkombinasikan Teknik Resampling Dan Algoritma Machine Learning Untuk Mengatasi Dataset Tak Seimbang”.

## Rumusan Masalah

Berdasarkan uraian pada latar belakang masalah di atas, dapat dikemukakan pertanyaan penelitian sebagai berikut:

1. Algoritma Machine Learning yang mana kah yang memiliki performa terbaik dalam mengklasifikasi dataset tak seimbang?
2. Teknik Resampling yang mana kah yang memiliki performa terbaik pada dataset tak seimbang?
3. Kombinasi algoritma Machine Learning dan teknik Resampling yang mana kah yang memiliki performa terbaik untuk mengatasi dataset tak seimbang?
4. Teknik resampling yang manakah yang lebih baik digunakan? Oversampling, undersampling, atau gabungan keduanya?

## Batasan Masalah

Batasan masalah pada penelitian ini adalah:

1. Dataset yang digunakan adalah dataset tak seimbang.
2. Dataset hanya memiliki dua kelas (biner).
3. Teknik resampling yang digunakan adalah Random Undersampling, Tomek Links, SBC, Random Oversampling, SMOTE, MSMOTE, Borderline-SMOTE, dan ADASYN.
4. Algoritma machine learning yang digunakan adalah Regresi Logistik, Decision Tree, Random Forest, Neural Network, Nearest Neighbor, dan Support Vector Machine.

## Tujuan Penelitian

Berdasarkan rumusan masalah, maka tujuan dari penelitian ini adalah:

1. Mengetahui algoritma machine learning yang memiliki performa terbaik dalam mengklasifikasi dataset tak seimbang.
2. Mengetahui teknik resampling yang memiliki performa terbaik pada dataset tak seimbang.
3. Mengetahui kombinasi algoritma machine learning dan teknik resampling yang memiliki performa terbaik dalam mengatasi dataset tak seimbang.
4. Mengetahui apakah oversampling, undersampling, atau gabungan keduanya yang lebih baik digunakan untuk masalah dataset tak seimbang.

## Manfaat Penelitian

Hasil penelitian ini diharapkan dapat bermanfaat:

1. Sebagai rujukan untuk mengatasi dataset tak seimbang yang sangat sering dijumpai.
2. Menjadi sumber informasi mengenai performa beberapa teknik resampling.
3. Menjadi sumber informasi mengenai performa dari beberapa algoritma machine learning.

## Organisasi Skripsi

# TINJAUAN PUSTAKA

Suatu deskripsi

## Landasan Teori

### ROC Curve & Confusion Matrix

ROC Curve (The Receiver Operating Characteristic Curve) adalah teknik untuk meringkas kinerja pengklasifikasi berdasarkan kesalahan positif (false positive) dan ke

### Cross Validation

### Teknik Resampling

#### Random Oversampling

Random oversampling, atau oversampling secara acak adalah teknik oversampling di mana anggota dari kelas minoritas dipilih secara acak dan diduplikasi ke dataset yang baru hingga tercapai keseimbangan [34]. Teknik ini juga biasa disebut teknik Bootstrap.

#### Synthetic Minority Oversampling Technique (SMOTE)

SMOTE atau Synthetic Minority Oversampling Technique adalah teknik oversampling terpopuler yang diproposalkan oleh Chawla [35] pada tahun 2002. Teknik ini membuat data tiruan atau sintetik berdasarkan tetangga terdekat dari sampel kelas minoritas.

#### Modified Synthetic Minority Oversampling Technique (MSMOTE)

#### Borderline – Synthetic Minority Oversampling Technique

#### Adaptive Synthetic (ADASYN)

#### Random Undersampling

#### Tomek Links

#### Sampling Based on Clustering (SBC)

### Algoritma Klasifikasi

#### Logistic Regression

#### Naive Bayes

#### Decision Tree

#### Random Forest

#### Neural Network

#### Nearest Neighbor

#### Support Vector Machine

## Penelitian Terkait

### Influence of Resampling on Accuracy of Imbalanced Classification

### Comparing oversampling techniques to handle the class imabalance problem: A customer churn prediction case study

### An approach for classification of highly imbalanced data using weighting and undersampling

### A study of the Behavior of Several Methods for Balancing Machine Learning Training Data

# METODE PENELITIAN

Suatu deskripsi

## Waktu dan Tempat

Penelitian ini dilaksanakan dari bulan Agustus 2019 sampai dengan bulan Oktober 2019. Lokasi penelitian dilakukan di Laboratorium Rekayasa Perangkat Lunak Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Hasanuddin

## Tahapan Penelitian

## Sumber Data

Data diambil dari Website resmi Kaggle (kaggle.com) dan UCI Machine Learning Repository (archive.ics.uci.edu/ml/). Data tersebut berupa tiga dataset berbeda, yaitu:

1. Credit Card Fraud Dataset (Kaggle), yang terdiri dari 30 kolom attribut dengan 1 kolom kelas, 284.807 baris.
2. Email Spam Dataset (UCI), yang terdiri dari 57 kolom atribut dengan 1 kolom kelas, 4.601 baris.

## Instrumen Penelitian

# HASIL DAN PEMBAHASAN

Suatu deskripsi

## asd

## ssd

## asdf

# KESIMPULAN DAN SARAN

## Kesimpulan

## Saran

Daftar Pustaka

|  |  |
| --- | --- |
| [1] | H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge & Data Engineering,* no. 9, pp. 1263-1284, 2008. |
| [2] | J. Desjardins, "How much data is generated each day?," World Economic Forum, 2019. |
| [3] | M. Chen, S. Mao and Y. & Liu, "Big Data: A survey," *Mobile networks and applications,* vol. 19, no. 2, pp. 171-209, 2014. |
| [4] | S. Lohr, "The age of big data," New York Times, New York, 2012. |
| [5] | N. Elgendy and A. Elragal, "Big data analytics: a literature review paper," *Industrial Conference on Data Mining,* pp. 214-227, 2014. |
| [6] | LinkedIn Economic Graph Team, "Linkedin 2018 Emerging Jobs Report," LinkedIn, 2018. |
| [7] | A. McAfee, E. Brynjolfsson, T. H. Davenport, D. J. Patil and D. Barton, "Big data: the management revolution," *Harvard business review,* vol. 90, no. 10, pp. 60-68, 2012. |
| [8] | V. Beal, "Unstructured Data," 2019. [Online]. Available: https://www.webopedia.com/TERM/U/unstructured\_data.html. [Accessed 20 6 2019]. |
| [9] | H. Baars and H. G. Kemper, "Management support with structured and unstructured data—an integrated business intelligence framework," *Information Systems Management,* vol. 25, no. 2, pp. 132-148, 2008. |
| [10] | G. Weglarz, "Two Worlds of Data - Unstructured and Structured," *DM Review,* vol. 14, pp. 19-23, 2004. |
| [11] | F. Berman, R. Rutenbar, B. Hailpern, H. Christensen, S. Davidson, D. Estrin, M. Franklin, M. Martonosi, P. Raghavan, V. Stodden and A. S. Szalay, "Realizing the Potential of Data Science," *Communications Of The Acm,* vol. 61, no. 4, pp. 67-72, 2018. |
| [12] | V. Dhar, "Data Science and Prediction," *Communications of the ACM,* vol. 56, no. 12, pp. 64-73, 2012. |
| [13] | F. Provost and T. Fawcett, "Data Science and its Relationship to Big Data and Data-Driven Decision Making," *Big data,* vol. 1, no. 1, pp. 51-59, 2013. |
| [14] | J. M. I. and M. T. M., "Machine learning: Trends, perspectives, and prospects," *Science,* vol. 349, no. 6245, pp. 255-260, 2015. |
| [15] | P. Domingos, "A Few Useful Things to Know about Machine Learning," pp. 78-87, 2011. |
| [16] | A. Ethem, Introduction to Machine Learning, MIT press, 2009. |
| [17] | C. M. Bishop, Neural Networks for Pattern Recognition, Oxford university press, 1995. |
| [18] | C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006. |
| [19] | F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Muller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and M. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research 12,* 2011. |
| [20] | M. Galar, A. Fernandez, E. Barrenechea, H. Bustince and F. Herrera, "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* vol. 42, no. 4, pp. 463-484, 2011. |
| [21] | S. Kotsiantis, D. Kanellopoulos and P. Pintelas, "Handling imbalanced datasets: A review," *GESTS International Transactions on Computer Science and Engineering,* vol. 30, no. 1, pp. 25-36, 2006. |
| [22] | A. Kumar and H. Sheshadri, "On the Classification of Imbalanced Datasets," *International Journal of Computer Applications (0975 – 8887),* vol. 44, no. 8, pp. 1-7, 2012. |
| [23] | S. Visa and A. Ralescu, "Issues in Mining Imbalanced Data Sets - A Review Paper," *Proceedings of the sixteen midwest artificial intelligence and cognitive science conference,* vol. 2005, pp. 67-73, April 2005. |
| [24] | F. Provost, "Machine Learning from Imbalanced Data Sets 101," *Proceedings of the AAAI’2000 workshop on imbalanced data sets,* vol. 68, pp. 1-3, July 2000. |
| [25] | M. M. Rahman and D. N. Davis, "Addressing the Class Imbalance Problem in Medical Datasets," *International Journal of Machine Learning and Computing,* vol. 3, no. 2, p. 224, 2013. |
| [26] | Statistics Solution, "Statistics Solution," 2016. [Online]. Available: https://www.statisticssolutions.com/sample-size-calculation-and-sample-size-justification-resampling/. [Accessed 14 August 2019]. |
| [27] | E. Burnaev, P. Erofeev and A. Papanov, "Influence of Resampling on Accuracy of Imbalanced Classification," *In Eighth International Conference on Machine Vision (ICMV 2015),* vol. 9875, p. 987521, 2015. |
| [28] | A. Anand, G. Pugalenthi, G. B. Fogel and P. N. Suganthan, "An approach for classification of highly imbalanced data using weighting and undersampling," *Amino acids,* vol. 39, no. 5, pp. 1385-1391, 2010. |
| [29] | A. More, "Survey of resampling techniques for improving classification performance in unbalanced datasets," *arXiv preprint arXiv:1608.06048,* 2016. |
| [30] | S. J. Yen and Y. S. Lee, "Under-Sampling Approaches for Improving Prediction of the Minority Class in an Imbalanced Dataset," *Intelligent Control and Automation,* pp. 731-740, 2006. |
| [31] | B. Diri and S. Albayrak, "Visualization and analysis of classifiers performance in multi-class medical data," *Expert Systems with Applications,* vol. 34, no. 1, pp. 628-634, 2008. |
| [32] | G. E. Batista, R. C. Prati and M. C. Monard, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," *ACM SIGKDD explorations newsletter,* vol. 6, no. 1, pp. 20-29, 2004. |
| [33] | A. Amin, S. Anwar, A. Adnan, M. Nawaz, N. Howard, J. Qadir and A. Hawalah, "Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study," *IEEE Access,* vol. 4, pp. 7940-7957, 2016. |
| [34] | C.-C. Chang and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology,* 2011. |
| [35] | J. Suykens and J. Vandewalle, "Least Squares Support Vector Machine Classifiers," *Neural Processing Letters 9,* 1999. |
| [36] | D. W. Hosmer and S. Lemeshow, Applied Logistic Regression (Second Edition), Canada: Wiley-Interscience Publication, 2000. |
| [37] | I. Rish, "An Empirical Study of the Naive Bayes Classifier," *IJCAI 2001 workshop on empirical methods in artificial intelligence,* pp. 41-46, 2001. |
| [38] | H. Zhang, "The Optimality of Naive Bayes," *American Association for Artificial Intelligence,* 2004. |
| [39] | D. D. Lewis, "Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval," *European conference on machine learning,* pp. 4-15, 1998. |
| [40] | L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, Classification and Regression Trees, Routledge: CRC Press LLC, 2017. |
| [41] | A. Liaw and M. Wiener, "Classification and regression by randomForest," *R news 2 no. 3,* pp. 18-22, 2002. |
| [42] | S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE transactions on systems, man, and cybernetics 21 no.3,* pp. 660-674, 1991. |
| [43] | M. Pal, "Random forest classifier for remote sensing classification," *International Journal of Remote Sensing 26 no.1,* pp. 217-222, 2005. |
| [44] | R. &. D. A. S. A. Díaz-Uriarte, "Gene selection and classification of microarray data using random forest," *BMC Bionformatics 7 no. 1,* p. 3, 2006. |
| [45] | J. M. Zurada, Introduction to artificial neural systems, St. Paul: West publishing company, 1992. |
| [46] | J. Kittler, M. Hater and R. P. Duin, "Combining classifiers," *Proceedings of 13th international conference on pattern recognition,* pp. 897-901, 1996. |
| [47] | B. Scholkopf and A. J. Smola, Learning with kernels: support vector machines, regularization, optimization, and beyond, MIT press, 2001. |
| [48] | Z. H. Zhou and X. Y. Liu, "Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem," *IEEE Transactions on Knowledge & Data Engineering,* no. 1, pp. 63-77, 2006. |
| [49] | S. Haykin, Neural Networks: A Comprehensive Foundation, Prentice Hall PTR, 1994. |
| [50] | G. P. Zhang, "Neural Networks for Classification: A Survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* vol. 30, no. 4, pp. 451-462, 2000. |
| [51] | N. V. Chawla, DATA MINING FOR IMBALANCED DATASETS: AN OVERVIEW, Boston: Springer, 2009, pp. 875-886. |
| [52] | A. Ali, S. M. Shamsuddin and A. L. Ralescu, "Classification with class imbalance problem: A Review," *Int. J. Advance Soft Compu. Appl,* vol. 7, no. 3, pp. 176-204, 2015. |
| [53] | S. M. A. Elrahman and A. Abraham, "A Review of Class Imbalance Problem," *Journal of Network and Innovative Computing,* pp. 332-340, 2013. |
| [54] | H. Guo and H. L. Viktor, "Learning from Imbalanced Data Sets with Boosting and Data Generation: The DataBoost-IM Approach," *ACM Sigkdd Explorations Newsletter,* vol. 6, no. 1, pp. 30-39, 2004. |
| [55] | D. L. Donoho and J. Tanner, "Precise Undersampling Theorems," *Proceedings of the IEEE,* vol. 98, no. 6, pp. 913-924, 2010. |
| [56] | C. Drummond and R. C. Holte, "C4.5, Class Imbalance, and Cost Sensitivity: Why Under-Sampling beats Over-Sampling," *Workshop on learning from imbalanced datasets II,* vol. 11, pp. 1-8, August 2003. |
| [57] | W. C. Lin, C. F. Tsai, Y. H. Hu and J. S. Jhang, "Clustering-based undersampling in class-imbalanced data," *Information Sciences,* vol. 409, pp. 17-26, 2017. |
| [58] | S. J. Yen and Y. S. Lee, "Cluster-based under-sampling approaches for imbalanced data distributions," *Expert Systems with Applications,* vol. 36, no. 3, pp. 5718-5727, 2009. |
| [59] | S. García and F. Herrera, "Evolutionary Undersampling for Classification with Imbalanced Datasets: Proposals and Taxonomy," *Evolutionary computation,* vol. 17, no. 3, pp. 275-306, 2009. |
| [60] | M. Kubat and S. Matwin, "Addressing the Curse of Imbalanced Training Sets: One-Sided Selection," *Icml,* vol. 97, pp. 179-186, July 1997. |
| [61] | X. Y. Liu, J. Wu and Z. H. Zhou, "Exploratory Undersampling for Class-Imbalance Learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics),* vol. 39, no. 2, pp. 539-550, 2008. |
| [62] | N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of artificial intelligence research,* vol. 16, pp. 321-357, 2002. |
| [63] | S. L. Y. M. L. &. H. Y. Hu, "MSMOTE: Improving Classification Performance When Training Data is Imbalanced," *2009 second international workshop on computer science and engineering,* vol. 2, pp. 13-17, October 2009. |
| [64] | H. Han, W. Y. Wang and B. H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," *International conference on intelligent computing,* pp. 878-887, August 2005. |
| [65] | H. He, Y. Bai, E. A. Garcia and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence),* pp. 1322-1328, June 2008. |
| [66] | IBM Corporation, "What is a data set?," IBM Corporation, [Online]. Available: https://www.ibm.com/support/knowledgecenter/zosbasics/com.ibm.zos.zconcepts/zconc\_datasetintro.htm. [Accessed 14 June 2019]. |
| [67] | H. Chen, R. H. Chiang and V. C. Storey, " Business intelligence and analytics: From big data to big impact," *MIS quarterly,* vol. 36, no. 4, 2012. |
| [68] | Y. Wang, L. Kung and T. A. Byrd, "Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations," *Technological Forecasting and Social Change,* vol. 126, pp. 3-13, 2018. |
| [69] | R. Blumberg and S. Atre, "The Problem with Unstructured Data," *DM Review,* vol. 13, no. 42-49, p. 62, 2003. |
| [70] | S. J. Russell and P. Norvig, Aritificial Intelligence: A Modern Approach, Pearson Education Limited, 2016. |

LAMPIRAN