**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

[Gambar 1 Kurva ROC 27](#_Toc22210354)

Persamaan

[Persamaan 1 sesuatu 14](#_Toc22210417)

Tabel

[Tabel 1 Confusion Matrix 25](#_Toc22210334)

Lampiran

[Lampiran 1 16](#_Toc22210463)

# PENDAHULUAN

## Latar Belakang

Dalam beberapa tahun terakhir, perkembangan yang pesat dalam sains dan teknologi telah berdampak pada pertumbuhan data mentah secara eksponensial (He & Garcia, Learning from Imbalanced Data, 2008). Berdasarkan World Economic Forum (Desjardins, 2019), 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 (Chen, Mao, & & Liu, 2014; Lohr, 2012).

Dari pertumbuhan data tersebut, kebutuhan untuk menganalisis data terus meningkat (Elgendy & Elragal, 2014). Menurut laporan dari LinkedIn pada tahun 2018, permintaan pekerjaan yang membutuhkan analisis data berkembang hingga 12 kali lipat lebih banyak dari tahun 2014 (LinkedIn Economic Graph Team, 2018). Ini disebabkan karena data merupakan kunci dari setiap organisasi, institusi atau perusahaan untuk mengambil keputusan yang lebih cermat dan efektif (McAfee, Brynjolfsson, Davenport, Patil, & Barton, 2012).

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 (Beal, 2019), di mana atribut data dapat dibedakan dengan jelas sehingga dapat langsung diproses oleh peralatan komputasi (Baars & Kemper, 2008). 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 (Weglarz, 2004). Menurut laporan Beal (Beal, 2019), 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 (Berman, et al., 2018; Dhar, 2012). Data Science melibatkan prinsip, proses, dan teknik untuk memahami fenomena melalui data (Provost & Fawcett, 2013). Data Science merupakan bidang yang sangat luas dan sedang dikembangkan (Provost & Fawcett, 2013), namun salah satu bidang khusus dari data science adalah *machine learning* yang merupakan paduan antar *computer science* (ilmu komputer) dan *statistics* (statistika) (I. & M., 2015). Machine learning membahas mengenai bagaimana membangun sistem komputer yang dapat belajar melalui pengalaman tanpa harus diprogram secara spesifik dan manual (I. & M., 2015) (Domingos, 2011).

Untuk menyelesaikan suatu masalah pada komputer, dibutuhkan algoritma. Namun untuk beberapa masalah, tidak ada algoritma yang memadai. Contohnya adalah bagaimana komputer mengenali angka dalam bentuk tulisan tangan (Bishop, Neural Networks for Pattern Recognition, 1995) dan mengklasifikasi suatu email menjadi spam atau bukan spam (Ethem, 2009). Dalam masalah seperti ini, seluruh objek dengan label atau kelas berbeda dikumpulkan dan diidentifikasi ciri-ciri 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, kelas, 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 (Bishop, Neural Networks for Pattern Recognition, 1995). 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, pembelajaran di mana mesin dilatih untuk membuat keputusan tertentu dengan cara *trial and error* disebut *reinforcement learning* (Bishop, Pattern Recognition and Machine Learning, 2006). Masing-masing jenis pembelajaran memiliki banyak algoritma yang telah dikembangkan dengan berbagai pendekatan yang berbeda-beda (I. & M., 2015). Berdasarkan dokumentasi dari *scikit-learn* (Pedregosa, et al., 2011), 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 (Galar, Fernandez, Barrenechea, Bustince, & Herrera, 2011). 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) (He & Garcia, Learning from Imbalanced Data, 2008). 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 (He & Garcia, Learning from Imbalanced Data, 2008). Sebagian besar data asli di dunia terdistribusi secara tak seimbang (He & Garcia, Learning from Imbalanced Data, 2008; Kotsiantis, Kanellopoulos, & Pintelas, 2006; Kumar & Sheshadri, 2012; Visa & Ralescu, 2005).

Pada umumnya, algoritma-algoritma machine learning, dalam hal ini pada masalah klasifikasi, bekerja dengan tujuan utama memaksimalkan akurasi (Provost F. , 2000). Hal ini sangat masuk akal, karena akurasi yang tinggi menjelaskan bahwa model algoritma tersebut melaksanakan tugasnya dengan baik, mengklasifikasikan kelas 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 akurasi yang akan diperoleh 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 (Visa & Ralescu, 2005; Rahman & Davis, 2013). 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 (Statistics Solution, 2016). Teknik resampling terdiri dari *oversampling*, yaitu mengambil sampel berulang kali dari kelas minoritas; dan undersampling, yaitu mengambil sampel secara acak dari kelas mayoritas (Burnaev, Erofeev, & Papanov, 2015). Kedua teknik ini dapat digunakan secara terpisah ataupun digabung (Burnaev, Erofeev, & Papanov, 2015; Anand, Pugalenthi, Fogel, & Suganthan, 2010; More, 2016; Yen & Lee, 2006). Masing-masing teknik resampling memiliki metode yang berbeda-beda, dengan performa yang berbeda pula (Burnaev, Erofeev, & Papanov, 2015; More, 2016). Begitu juga dengan algoritma klasifikasi pada machine learning, terdapat banyak metode yang berbeda dengan performa yang berbeda pula (Diri & Albayrak, 2008).

Dalam beberapa penelitian terkait (More, 2016; Batista, Prati, & Monard, 2004; Amin, et al., 2016; Burnaev, Erofeev, & Papanov, 2015), 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. (Amin, et al., 2016) hanya meneliti teknik oversampling, (Burnaev, Erofeev, & Papanov, 2015; More, 2016; Batista, Prati, & Monard, 2004) meneliti teknik oversampling dan undersampling namun hanya menggunakan satu algoritma machine learning, sedangkan Diri (Diri & Albayrak, 2008) 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. Kombinasi algoritma Machine Learning dan teknik Resampling yang mana kah yang memiliki performa terbaik untuk mengatasi dataset tak seimbang?

## Batasan Masalah

Batasan masalah pada penelitian ini adalah:

1. Merupakan masalah klasifikasi.
2. Dataset yang digunakan adalah dataset tak seimbang.
3. Dataset hanya memiliki dua kelas (biner).
4. Atribut-atribut dataset merupakan tipe numerik.
5. Teknik resampling yang digunakan adalah Random Undersampling, Tomek Links, Random Oversampling, SMOTE, Borderline-SMOTE, dan ADASYN.
6. Algoritma machine learning yang digunakan adalah Regresi Logistik, Decision Tree, Naive Bayes, 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 sering dijumpai.
2. Menjadi sumber informasi mengenai performa beberapa teknik resampling.
3. Menjadi sumber informasi mengenai performa dari beberapa algoritma machine learning.

# TINJAUAN PUSTAKA

## Machine Learning

Machine learning adalah memprogram komputer untuk mengoptimalkan suatu ukuran kinerja menggunakan sampel data atau berdasarkan pengalaman (Ethem, 2009). Machine learning menggunakan suatu algoritma untuk menganalisis data.

Dalam pembelajaran yang terawasi atau *supervised learning*, pengklasifikasi akan diberikan suatu input tertentu dan menghubungkannya dengan suatu output. 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*. Dalam pembelajaran tanpa pengawasan atau *unsupervised learning*, pengklasifikasi diberi input dan dibiarkan sendiri untuk menemukan pola pada data tersebut. Kasus *unsupervised learning* di mana tujuannya adalah mengelompokkan observasi-observasi yang mirip disebut *clustering*, jika menentukan distribusi data pada input disebut *estimasi kepadatan*. Dalam *reinforcement learning*, sistem komputer menerima input secara terus menerus dan mencoba memilih keputusan-keputusan yang paling optimal berdasarkan kondisi lingkungannya.

Masing-masing jenis pembelajaran memiliki banyak algoritma yang telah dikembangkan dengan berbagai pendekatan yang berbeda-beda (I. & M., 2015).

## Dataset

Dataset adalah kumpulan data yang berbentuk tabel, di mana setiap kolomnya merepresentasikan suatu ciri-ciri, atribut atau fitur. Setiap barisnya menyatakan observasi suatu individu, record atau sampel (Snijders, Matzat, & Reips, 2012). Suatu dataset biasanya memiliki satu kolom tambahan yang merepresentasikan kelas dari observasi tersebut, kolom ini disebut kolom kelas. Kolom kelas ini juga disebut sebagai variabel dependen terhadap variabel-variabel independen yang merupakan ciri-ciri (atribut) dari suatu observasi tertentu.

Dalam machine learning dikenal istilah dataset yang tak seimbang. Istilah ini berlaku ketika kelas dari dataset tersebut bersifat kategorik diskrit. Dataset yang tak seimbang (*imbalanced dataset*) adalah dataset yang frekuensi kejadian dari kelas tertentu sangat jauh berbeda dengan kelas yang lain. Contohnya seperti suatu dataset dengan jumlah pasien yang berkelas “diabetes” jumlahnya jauh lebih sedikit dibanding pasien yang “tidak diabetes”.

Masalah ketidakseimbangan ini akan memberi bias terhadap performa pengklasifikasi sebab jumlah sampel pada kelas tertentu tidak dapat memberi informasi yang cukup kepada pengklasifikasi berdasarkan ciri-ciri yang diberikan (Bishop, Pattern Recognition and Machine Learning, 2006; Ethem, 2009; Domingos, 2011).

## ROC Curve

Di dalam machine learning, mengukur kinerja atau performa dari suatu model adalah hal yang esensial. Model yang diperoleh dari pelatihan melalui data training perlu diuji melalui data testing. Kinerja diukur berdasarkan seberapa baik model tersebut memprediksi dengan benar data yang ada.

Pada klasifikasi biner, kelas positif yang berhasil diprediksi dengan benar disebut true positive, jika kelas positif tersebut diprediksi negatif (salah) disebut false negative. Kelas negatif yang berhasil diprediksi negatif (benar) disebut true negative, dan kelas negatif yang diprediksi positif disebut false positive. Jumlah dari kasus-kasus tersebut direpresentasikan dalam suatu tabel kontingensi yang disebut *confusion matrix* (Swets, 1988)*.*

|  |  |  |  |
| --- | --- | --- | --- |
|  | Kelas asli | | |
| Hasil prediksi |  | Positif | Negatif |
| Positif | TP | FP |
| Negatif | FN | TN |

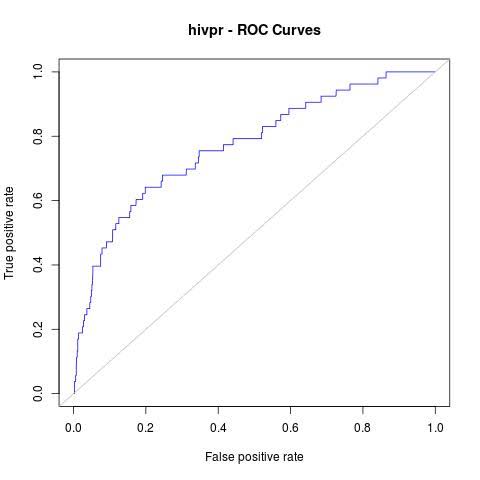
Tabel 1 Confusion Matrix

Akurasi adalah ukuran kinerja yang menunjukkan seberapa baik suatu pengklasifikasi dalam mengklasifikasikan seluruh data. Akurasi adalah rasio antara observasi yang diklasifikasikan secara benar dengan total observasi:

Presisi adalah ukuran kinerja yang menunjukkan seberapa besar kebenaran suatu pengklasfikasi dari seluruh kelas positif yang diprediksi. Presisi adalah rasio antara jumlah kelas positif yang diklasifikasikan secara benar dengan jumlah observasi yang diklasifikasikan positif:

Recall atau sensitivitas adalah ukuran kinerja yang menunjukkan seberapa baik suatu pengklasifikasi dalam mengklasifikasikan kelas positif. Recall adalah rasio antara jumlah observasi positif yang diklasifikasikan secara benar dengan jumlah observasi positif asli:

ROC (Receiver Operating Characteristic) Curve atau Kurva ROC adalah teknik yang diproposalkan oleh Swets (Swets, 1988) untuk meringkas kinerja pengklasifikasi berdasarkan rasio positif asli (true positive) dengan positif palsu (false positive). Kurva ROC menunjukkan kinerja suatu model berdasarkan threshold atau nilai ambang batas tertentu yang digunakan pada model tersebut.



Gambar 1 Kurva ROC

Nilai AUC (*Area Under Curve*) menunjukkan nilai peluang suatu model memprediksi kelas positif sebagai positif dan kelas negatif sebagai negatif, atau seberapa baik suatu model membedakan antara kelas positif dan negatif. Semakin tinggi nilai AUC, maka semakin baik model tersebut dalam membedakan atara kelas positif dan negatif. (Swets, 1988; LeDell, Petersen, & Laan, 2015; Bradley, 1997). Nilai AUC 1 menunjukkan bahwa model tersebut dapat dengan sempurna membedakan antara kelas positif dengan kelas negaitf, dan nilai AUC 0,5 menunjukkan bahwa model tersebut sama sekali tidak dapat membedakan antara kelas positif dan negatif (Bradley, 1997).

## Jarak Euklid

Jarak Euklid adalah jarak antara suatu vektor ke suatu vektor pada ruang euklid berdimensi :

Pada machine learning, jarak Euklid digunakan untuk menghitung jarak antar dua observasi berdasarkan vektor fitur yang bersifat kontinu (Howard, 2013).

## 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 (Liu, 2004). Data minoritas tersebut dapat diduplikasi beberapa kali. Teknik ini biasanya menyebabkan overfitting pada model (Amin, et al., 2016; Liu, 2004).

### Synthetic Minority Oversampling Technique (SMOTE)

SMOTE atau Synthetic Minority Oversampling Technique adalah teknik oversampling terpopuler yang diproposalkan oleh Chawla (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) pada tahun 2002. Teknik ini membuat data tiruan atau sintetik berdasarkan tetangga-tetangga terdekat dari sampel kelas minoritas.

Teknik ini dimulai dengan menentukan data yang akan dibuat untuk setiap data pada kelas minoritas dalam dataset . Kemudian untuk setiap data kelas minoritas , pilih tetangga secara acak dari tetangga terdekat data tersebut di mana dengan adalah jumlah data pada kelas minoritas dan , di mana adalah fitur-fitur pada . Lalu untuk setiap fitur pada , hitung jarak antara dengan setiap tetangga tersebut:

di mana adalah bilangan acak dari tetangga terdekat dengan T adalah salah satu tetangga terdekat . Kemudian suatu bilangan acak ditentukan. Data tiruan dibuat berdasarkan:

di mana bersifat inkremental dan adalah data kelas minoritas tiruan.

Teknik ini membuat data sintetik pada suatu titik dari jarak antara setiap fitur dari dengan (Chawla, Bowyer, Hall, & Kegelmeyer, 2002).

### Borderline – Synthetic Minority Oversampling Technique (Borderline-SMOTE)

Terinspirasi oleh SMOTE, Han memproposalkan teknik oversampling baru yang menyerupai SMOTE. Teknik Han tersebut membuat data tiruan di sekitar data kelas minoritas yang berada di *borderline* atau perbatasan antara kelas mayoritas dan kelas minoritas saja, di mana SMOTE membuat data tiruan pada setiap sampel kelas minoritas (Han, Wang, & Mao, 2005).

Misalkan untuk data training terdapat kelas minoritas dan kelas mayoritas , diekspresikan:

di mana dan adalah jumlah sampel kelas minoritas dan jumlah sampel kelas mayoritas. Borderline-SMOTE bekerja sebagai berikut.

Untuk setiap pada kelas minoritas , hitung tetangga terdekat dari seluruh data training . Jumlah kelas mayoritas dari tetangga terdekat disimbolkan .

Jika , atau seluruh tetangga terdekat dari adalah sampel kelas mayoritas, maka dinyatakan sebagai noise dan bukan merupakan sampel perbatasan. Jika , atau tetangga terdekat lebih banyak merupakan sampel kelas mayoritas dibanding minoritas, maka dianggap sampel perbatasan dan dimasukkan ke dalam himpunan sampel perbatasan . Jika , maka dianggap aman dan bukan merupakan sampel perbatasan.

Untuk himpunan sampel perbatasan , di mana . Maka:

di mana adalah jumlah sampel perbatasan. Untuk setiap sampel perbatasan di B, tentukan tetangga terdekat dari .

Data tiruan berkelas minoritas dibuat sebanyak . Untuk setiap , Pilih sebanyak tetangga terdekat secara acak dari , lalu hitung jarak antara dengan tetangga terdekatnya dari di mana:

adalah jarak antara dengan salah satu tetangga terdekatnya. Kemudian suatu bilangan acak ditentukan dan data tiruan sebanyak dari dibuat.

Proses di atas dilakukan untuk setiap , maka data tiruan akan dibuat sebanyak kali (Han, Wang, & Mao, 2005).

### Adaptive Synthetic (ADASYN)

Adaptive Synthetic adalah teknik oversampling yang diproposalkan oleh He di mana data tiruan dibuat berdasarkan tingkat kesulitan suatu sampel kelas minoritas untuk dipelajari. Setiap sampel kelas minoritas memiliki alokasi data tiruan sesuai dengan banyaknya tetangga sampel kelas mayoritas dari tetangga terdekat sampel kelas minoritas tersebut (He, Bai, Garcia, & Li, 2008).

Misalkan untuk data training D dengan m sampel , , di mana adalah instansi dari dimensi dari ruang fitur dan adalah label kelas dari . Didefinisikan dan sebagai jumlah sampel kelas minoritas, dan jumlah sampel kelas mayoritas secara berurutan. Maka dan .

Tentukan derajat ketidakseimbangan :

di mana . Jika di mana adalah nilai maksimum toleransi derajat ketidakseimbangan, maka hitung jumlah data kelas minoritas tiruan:

di mana adalah parameter yang menunjukkan rasio ketidakseimbangan yang diinginkan setelah data tiruan dibuat. menunjukkan dataset akan seimbang sepenuhnya.

Untuk setiap sampel di mana adalah kelas minoritas, tentukan tetangga terdekat dengan jarak Euklid pada dimensi ruang fitur. Kemudian hitung rasio , yang didefinisikan sebagai:’

di mana adalah jumlah sampel pada tetangga terdekat yang merupakan sampel kelas mayoritas, maka .

Kemudian normalisasikan menjadi:

agar menjadi distribusi kepadatan .

Hitung data tiruan yang akan dibuat untuk setiap sampel :

Untuk setiap sampel , buat data tiruan dengan persamaan:

di mana adalah selisih vektor pada dimensi dan adalah bilangan acak .

### Random Undersampling

Random undersampling, atau undersampling secara acak adalah teknik undersampling di mana anggota dari kelas mayoritas dipilih secara acak dan dihapus dari dataset training hingga tercapai keseimbangan. Kekurangan dari teknik ini adalah tidak ada cara untuk mengatur informasi apa saja yang dihilangkan dari dataset tersebut, informasi yang berguna bisa saja hilang (Amin, et al., 2016; Liu, 2004; Yen & Lee, 2006; More, 2016).

### Tomek Links

Tomek Link atau tautan Tomek adalah metode undersampling yang diproposalkan oleh Tomek (Tomek, 1976) untuk memodifikasi CNN (Condensed Nearest Neighbor). Teknik ini menghapus sampel kelas mayoritas jika dan hanya tetangga kelas mayoritas tersebut berasal dari kelas berbeda dan merupakan tetangga terdekat satu sama lain. Jika dan adalah sampel dengan kelas berbeda maka Tomek Link didefinisikan untuk setiap sampel :

di mana fungsi adalah jarak Euklid (4).

## Algoritma Klasifikasi Machine Learning

### Logistic Regression

Regresi logistik adalah algoritma klasifikasi yang menggunakan kelas untuk membangun dan menggunakan model regresi logistik multinomial tunggal dengan *estimator* tunggal. Regresi logistik menyatakan probabilitas kelas (antara 0 dan 1) tergantung pada jarak dari batas, dengan sutu pendekatan. Hasil regresi logistik adalah suatu bilangan antara 0 dan 1 yang menyatakan probabilitas kelas.

Nilai atribut pada data observasi dari regresi logistik dapat berupa nominal, ordinal, interval atau skala rasio, sedangkan untuk atribut kelas harus merupakan kelas biner. Hubungan antara atribut dengan kelas bersifat non-linear. Distribusi data tidak tersebar dalam bentuk distribusi Gauss, melainkan distribusi Bernoulli, yang dikarenakan kelasnya berbentuk biner.

Pada analisis regresi logistik, hubungan antara suatu kejadian berdasarkan atribut atau vektor fiturnya diekspresikan:

di mana adalah peluang terjadinya suatu kejadian. Peluang tersebut bernilai nol hingga satu pada kurva berbentuk S dan adalah kombinasi linear dari vektor fitur kejadian tersebut:

di mana adalah nilai *intercept* dari model, dan adalah koefisien kemiringan atau gradien dari model regresi logistik tersebut, dan adalah nilai atribut ke- dari vektor fitur kejadian tersebut (Tsangaratos & Ilia, 2016; Hosmer & Lemeshow, 2000; F.Y., et al., 2017).

### Naive Bayes

Naive Bayes adalah algoritma klasifikasi yang memprediksi kelas suatu data berdasarkan vektor fitur atau atribut-atributnya. Pembelajaran pengklasifikasi ini mengasumsikan bahwa setiap fitur adalah kelas yang independen, yaitu:

di mana adalah fitur vektor dan adalah kelas (Rish I. , 2011).

Berdasarkan aturan Bayes, peluang sampel berlabel suatu kelas biner adalah:

dan E diklasifikasikan jika dan hanya jika:

di mana adalah pengklasifikasi Bayes. Asumsikan seluruh atribut bersifat independen terhadap variabel kelas, yaitu:

maka hasil pengklasifikasinya adalah:

disebut pengklasifikasi naive Bayes (Lewis, 1998; Rish I. , 2001; Zhang, 2004).

### Decision Tree

Decision Tree atau pohon keputusan adalah pengklasifikasi yang menentukan output kelas dari suatu sampel berdasarkan keputusan yang diambil dari setiap nilai atribut dari sampel tersebut. Pohon keputusan adalah suatu fungsi Boolean di mana inputnya adalah suatu sampel dengan vektor fitur dan outputnya adalah 0 atau 1. Di pohon keputusan, setiap node pohon yang bukan node daun adalah suatu uji atribut atau suatu ekspresi boolean, setiap node daun adalah nilai Boolean, dan setiap cabang mewakili salah satu nilai yang mungkin dari atribut yang diuji. Ada beberapa metode pohon keputusan seperti ID3, C4.5, dan CART.

Algoritma ID3 menggunakan teori entropi informasi yang menghitung entropi dari masing-masing atribut. Kemudian menghitung Information Gain dari suatu atribut berdasarkan entropinya terhadap target kelas. Di mana entropi adalah ukuran ketidapastian suatu atribut yang terkait dengan suatu kejadian. Semakin kecil entropi dari suatu atribut, semakin tinggi kemurnian informasi yang ada. Semakin besar entropi dari suatu atribut, semakin besar ketidakpastian informasi tersebut. Information Gain adalah jumlah ketidakpastian informasi yang berkurang berdasarkan informasi yang diterima terkait dengan suatu atribut.

Misalkan adalah data Training dengan jumlah sampel , dengan jumlah nilai atribut kelas yang berbeda dari . Misalkan adalah jumlah sampel pada . Informasi yang dibutuhkan dari suatu sampel klasifikasi adalah:

di mana adalah peluang suatu sampel berasal dari kelas yang diperoleh dari . Suatu himpunan atribut memiliki nilai berbeda . Bagi menjadi subset , di mana memiliki beberapa sampel di yang memiliki nilai di .

Misalkan adalah jumlah sampel di kelas dari subset , entropi subset yang dibagi oleh adalah:

Berdasarkan persamaan di atas, informasi yang dibutuhkan dari suatu subset dihitung berdasarkan:

Maka Information Gain dari A dapat dihitung dengan:

Algoritma ID3 membangun decision tree berdasarkan urutan atribut yang memiliki information gain tertinggi sebagai root node. Node tersebut kemudian bercabang sesuai dengan jumlah nilai berbeda pada atribut tersebut. Cabang dengan entropi nol adalah node daun, dan cabang dengan entropi yang lebih dari nol membutuhkan pecabangan lebih lanjut. Proses tersebut dilakukan secara rekursif pada semua node yang bukan merupakan node daun hingga seluruh data terklasifikasi (F.Y., et al., 2017; Quinlan, 1986; Dai, Zhang, & Wu, 2016; Vafeiadis, Diamantaras, Sarigiannidis, & Chatzisavvas, 55).

### Support Vector Machine

Support Vector Machine adalah algoritma pengklasifikasi yang membangun suatu hyperplane yang memisahkan data-data dengan kelas berbeda dengan margin sebesar mungkin.

Misalkan suatu data training pada sampel dipisahkan menjadi dua kelas, di mana adalah vektor fitur dan adalah label kelasnya. Jika diasumsikan dua kelas tersebut dapat dipisahkan dengan hyperplane pada suatu bidang , dan informasi mengenai distribusi datanya tidak diketahui, maka hyperplane yang optimal adalah hyperplane yang memaksimalkan margin. Nilai optimal dari dan dapat dicari menggunakan fungsi Lagrange :

di mana dan diperoleh dari algoritma pembelajaran SVC. di mana nilai tidak nol adalah “support vector”. Untuk adalah inner product atau hasil kali dalam antara dua observasi dan yang diekspresikan:

yang disebut kernel Linear.

Metode SVM yang menggunakan kernel linear dapat memisahkan kelas-kelas secara linear. Namun secara praktis dataset yang dijumpai sering tidak dapat dipisahkan secara linear. Kernel non-linear lebih fleksibel dikarenakan kernel ini memetakan variabel dan ke dimensi yang lebih tinggi. Salah satu kernel non-linear yang populer digunakan adalah kernel radial atau kernel Gauss:

di mana adalah parameter tuning tambahan. Ketika suatu sampel test terletak jauh dari sampel training, eksponensial tersebut akan menjadi negatif, dan mendekati nol. Pada kasus ini, sampel training yang terletak jauh dari sampel test hampir tidak memiliki efek atau kontribusi kepada keputusan yang digunakan untuk mengklasifikasikan sampel test. Parameter mempengaruhi seberapa jauh letak suatu observasi agar dapat berkontribusi terhadap keputusan klasifikasi. Kernel ini lebih fleksibel dari kernel linear (Guenther & Schonlau, 2016; Schuldt, Laptev, & Caputo, 2004; Vapnik & Cortes, 1995).

### Multilayer Perceptron (MLP)

Artificial Neural Network (ANN) atau Jaringan Saraf Tiruan adalah salah satu algoritma machine learning yang mencari suatu fungsi yang tidak diketahui berdasarkan input vektor ke suatu output :

Pada tahap pelatihan, fungsi dioptimasikan sedemikian sehingga hasil yang didapat untuk setiap vektor X yang diberikan sedekat mungkin dengan nilai .

Istilah neuron atau node digunakan pada ANN untuk menandakan suatu atribut (pada layer input), suatu jenis output (pada layer output) atau suatu ekspresi matematika (pada layer tersembunyi). Neuron atau node ini masing-masing memiliki *weight* (bobot) yang berbeda-beda yang mempengaruhi nilai output dari node tersebut berdasarkan suatu fungsi aktivasi. Untuk node pada layer non-output, output node tersebut akan dikirim ke seluruh node pada layer selanjutnya. Koneksi atau sambungan antara node memiliki nilai bias.

Multilayer Perceptron (MLP) adalah salah satu jenis Artificial Neural Network yang memiliki layer tesembunyi atau hidden layer. MLP terdiri dari setidaknya 3 layer, yaitu input layer, hidden layer dan output layer. Secara praktis, MLP biasanya memiliki lebih dari satu hidden layer.

MLP dapat diekspresikan sebagai berikut:

di mana adalah vektor input yang diaugmentasikan dengan 1, atau , dan adalah vektor bobot dari node tersembunyi ke-, dan adalah bobot untuk node output, dan adalah output jaringannya. Fungsi adalah fungsi aktivasi pada node tersembunyi (Ahmed, Atiya, Gayar, & El-Shishiny, 2010; Zurada, 1992; Haykin, 1994; Zhou & Liu, 2006; Bishop, Pattern Recognition and Machine Learning, 2006; Bishop, Neural Networks for Pattern Recognition, 1995; MULTILAYER, 1998).

Fungsi aktivasi adalah fungsi yang memberi output dari suatu node berdasarkan input yang diberikan.

Fungsi aktivasi Threshold atau Unit Step adalah fungsi aktivasi yang memberi nilai 0 atau 1 sesuai dengan threshold atau nilai batas yang ditentukan:

Fungsi aktivasi Sigmoid logistik adalah fungsi aktivasi yang memberi nilai antara 0 dan 1:

Fungsi aktivasi Sigmoid tangensial adalah fungsi aktivasi yang memberi nilai antara

-1 dan 1

### K-Nearest Neighbor

K-Nearest Neighbor adalah algoritma prediksi non-parametrik di mana hasil prediksi kelas dari suatu titik didasarkan oleh mayoritas kelas dari tetangga terdekatnya. Diberikan suatu titik, hitung jarak Euklid antara titik tersebut dengan semua titik pada data training. Kemudian pilih tetangga terdekat berdasarkan jarak Euklid tersebut, prediksi kelas dari titik tersebut adalah modus kelas dari tetangga terdekatnya.

Untuk suatu sampel dengan tetangga terdekat , prediksi kelas ditentukan oleh:

di mana adalah label kelas dari (Ahmed, Atiya, Gayar, & El-Shishiny, 2010; F.Y., et al., 2017).

# METODE PENELITIAN

## 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

Untuk menyelesaikan penelitian ini, peneliti akan melewati beberapa tahap penelitian, yaitu: Pra-penelitian, eksplorasi data, model *tuning and fitting*, dan analisis hasil.

Pada tahap pra-penelitian, peneliti menentukan tema penelitian, masalah yang akan diteliti, mengumpulkan sumber referensi atau literatur seperti jurnal dan buku yang mendukung dalam penelitian, dan menentukan metode yang digunakan beserta batasan masalahnya. Kemudian peneliti mencari data yang sesuai dengan tema penelitian sebagai objek penelitian.

Pada tahap eksplorasi data, peneliti mencoba menguraikan karakteristik-karakteristik data yang digunakan untuk menentukan fungsi-fungsi tertentu yang akan digunakan. Setiap dataset akan memberikan parameter-parameter optimal yang berbeda terhadap algoritma klasifikasi machine learning.

Pada tahap model *tuning and fitting*, peneliti akan mencari parameter-parameter terbaik untuk model yang akan digunakan berdasarkan hasil eksplorasi data dan *trial and error* untuk mendapatkan hasil terbaik. Tuning juga dilakukan terhadap beberapa teknik resampling yang membutuhkan parameter. Kemudian model akan memberi hasil prediksi yang akan dianalisis pada tahap selanjutnya.

Pada tahap analisis hasil, peneliti akan merangkum hasil yang diperoleh dari metode-metode yang digunakan ke dalam bentuk tabel dan diagram, kemudian menyimpulkan hasilnya sebagai output dari penelitian ini.

## Deskripsi Data

Data diambil dari Website resmi Kaggle (kaggle.com), UCI Machine Learning Repository (archive.ics.uci.edu/ml/) dan KEEL (sci2s.ugr.es/keel/imbalanced.php). Data tersebut berupa tiga dataset, yaitu:

1. Credit Card Fraud Dataset (Kaggle), yang terdiri dari 30 kolom attribut dengan 1 kolom kelas, 284.807 baris. 284.315 jumlah sampel kelas mayoritas dan 492 jumlah sampel kelas minoritas dengan *imbalanced ratio* sebesar 577:1. Dataset ini merupakan dataset terpopuler di Kaggle sebab jumlah data yang besar dengan imbalanced ratio yang sangat tinggi.
2. Spambase Dataset (UCI), yang terdiri dari 57 kolom atribut dengan 1 kolom kelas, 4.601 baris. 2788 jumlah sampel kelas mayoritas dan 1813 jumlah sampel kelas minoritas dengan *imbalanced ratio* sebesar 1,5:1
3. Image Segmentation Dataset (KEEL), yang terdiri dari 19 kolom atribut dengan 1 kolom kelas, 2308 baris. 1962 jumlah sampel kelas mayoritas dan 346 jumlah sampel kelas minoritas dengan *imbalanced ratio* sebesar 6:1.

Seluruh dataset hanya memiliki atribut kontinu dengan label kelas biner yang sesuai dengan tema penelitian dan metode-metode yang digunakan.

# HASIL DAN PEMBAHASAN

## Data Preprocessing

### Spambase Dataset

## ssd

## asdf

# KESIMPULAN DAN SARAN

## Kesimpulan

## Saran

Daftar Pustaka

Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2015). Classification with class imbalance problem: A Review. *Int. J. Advance Soft Compu. Appl, 7*(3), 176-204.

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., & Hawalah, A. (2016). Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study. *IEEE Access, 4*, 7940-7957.

Anand, A., Pugalenthi, G., Fogel, G. B., & Suganthan, P. N. (2010). An approach for classification of highly imbalanced data using weighting and undersampling. *Amino acids, 39*(5), 1385-1391.

Baars, H., & Kemper, H. G. (2008). Management support with structured and unstructured data—an integrated business intelligence framework. *Information Systems Management, 25*(2), 132-148.

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *ACM SIGKDD explorations newsletter, 6*(1), 20-29.

Beal, V. (2019). *Unstructured Data.* Retrieved 6 20, 2019, from https://www.webopedia.com/TERM/U/unstructured\_data.html

Berman, F., Rutenbar, R., Hailpern, B., Christensen, H., Davidson, S., Estrin, D., . . . Szalay, A. S. (2018). Realizing the Potential of Data Science. *Communications Of The Acm, 61*(4), 67-72.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition.* Oxford university press.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer.

Blumberg, R., & Atre, S. (2003). The Problem with Unstructured Data. *DM Review, 13*(42-49), 62.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification and Regression Trees.* Routledge: CRC Press LLC.

Burnaev, E., Erofeev, P., & Papanov, A. (2015). Influence of Resampling on Accuracy of Imbalanced Classification. *In Eighth International Conference on Machine Vision (ICMV 2015), 9875*, 987521.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*.

Chawla, N. V. (2009). *DATA MINING FOR IMBALANCED DATASETS: AN OVERVIEW.* Boston: Springer.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of artificial intelligence research, 16*, 321-357.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS quarterly, 36*(4).

Chen, M., Mao, S., & & Liu, Y. (2014). Big Data: A survey. *Mobile networks and applications, 19*(2), 171-209.

Desjardins, J. (2019, April 17). *How much data is generated each day?* World Economic Forum. Retrieved June 14, 2019, from https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/

Dhar, V. (2012). Data Science and Prediction. *Communications of the ACM, 56*(12), 64-73.

Díaz-Uriarte, R. &. (2006). Gene selection and classification of microarray data using random forest. *BMC Bionformatics 7 no. 1*, 3.

Diri, B., & Albayrak, S. (2008). Visualization and analysis of classifiers performance in multi-class medical data. *Expert Systems with Applications, 34*(1), 628-634.

Domingos, P. (2011). A Few Useful Things to Know about Machine Learning. 78-87.

Donoho, D. L., & Tanner, J. (2010). Precise Undersampling Theorems. *Proceedings of the IEEE, 98*(6), 913-924.

Drummond, C., & Holte, R. C. (2003, August). C4.5, Class Imbalance, and Cost Sensitivity: Why Under-Sampling beats Over-Sampling. *Workshop on learning from imbalanced datasets II, 11*, 1-8.

Elgendy, N., & Elragal, A. (2014). Big data analytics: a literature review paper. *Industrial Conference on Data Mining*, 214-227.

Elrahman, S. M., & Abraham, A. (2013). A Review of Class Imbalance Problem. *Journal of Network and Innovative Computing*, 332-340.

Ethem, A. (2009). *Introduction to Machine Learning.* MIT press.

Fernandez, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research, 61*, 863-905.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2011). 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), 42*(4), 463-484.

García, S., & Herrera, F. (2009). Evolutionary Undersampling for Classification with Imbalanced Datasets: Proposals and Taxonomy. *Evolutionary computation, 17*(3), 275-306.

Guo, H., & Viktor, H. L. (2004). Learning from Imbalanced Data Sets with Boosting and Data Generation: The DataBoost-IM Approach. *ACM Sigkdd Explorations Newsletter, 6*(1), 30-39.

Han, H., Wang, W. Y., & Mao, B. H. (2005, August). Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. *International conference on intelligent computing*, 878-887.

Haykin, S. (1994). *Neural Networks: A Comprehensive Foundation.* Prentice Hall PTR.

He, H., & Garcia, E. A. (2008). Learning from Imbalanced Data. *IEEE Transactions on Knowledge & Data Engineering*(9), 1263-1284.

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008, June). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 1322-1328.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression (Second Edition).* Canada: Wiley-Interscience Publication.

Hu, S. L. (2009, October). MSMOTE: Improving Classification Performance When Training Data is Imbalanced. *2009 second international workshop on computer science and engineering, 2*, 13-17.

I., J. M., & M., M. T. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255-260.

IBM Corporation. (n.d.). *What is a data set?* (IBM Corporation) Retrieved June 14, 2019, from https://www.ibm.com/support/knowledgecenter/zosbasics/com.ibm.zos.zconcepts/zconc\_datasetintro.htm

Kittler, J., Hater, M., & Duin, R. P. (1996). Combining classifiers. *Proceedings of 13th international conference on pattern recognition*, 897-901.

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering, 30*(1), 25-36.

Kubat, M., & Matwin, S. (1997, July). Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. *Icml, 97*, 179-186.

Kumar, A., & Sheshadri, H. (2012). On the Classification of Imbalanced Datasets. *International Journal of Computer Applications (0975 – 8887), 44*(8), 1-7.

Lewis, D. D. (1998). Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. *European conference on machine learning*, 4-15.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news 2 no. 3*, 18-22.

Lin, W. C., Tsai, C. F., Hu, Y. H., & Jhang, J. S. (2017). Clustering-based undersampling in class-imbalanced data. *Information Sciences, 409*, 17-26.

LinkedIn Economic Graph Team. (2018). *Linkedin 2018 Emerging Jobs Report.* LinkedIn.

Liu, A. Y.-c. (2004). The effect of oversampling and undersampling on classifying imbalanced text datasets. *The University of Texas at Austin*.

Liu, X. Y., Wu, J., & Zhou, Z. H. (2008). Exploratory Undersampling for Class-Imbalance Learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39*(2), 539-550.

Lohr, S. (2012). *The age of big data.* New York: New York Times.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: the management revolution. *Harvard business review, 90*(10), 60-68.

More, A. (2016). Survey of resampling techniques for improving classification performance in unbalanced datasets. *arXiv preprint arXiv:1608.06048*.

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing 26 no.1*, 217-222.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, M. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research 12*.

PENG, C.-Y. J., LEE, K. L., & INGERSOLL, G. M. (2002). An introduction to logistic regression analysis and reporting. *The journal of educational research, 96*(1), 3-14.

Provost, F. (2000, July). Machine Learning from Imbalanced Data Sets 101. *Proceedings of the AAAI’2000 workshop on imbalanced data sets, 68*, 1-3.

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big data, 1*(1), 51-59.

Rahman, M. M., & Davis, D. N. (2013). Addressing the Class Imbalance Problem in Medical Datasets. *International Journal of Machine Learning and Computing, 3*(2), 224.

Rish, I. (2001). An Empirical Study of the Naive Bayes Classifier. *IJCAI 2001 workshop on empirical methods in artificial intelligence*, 41-46.

Russell, S. J., & Norvig, P. (2016). *Aritificial Intelligence: A Modern Approach.* Malaysia: Pearson Education Limited.

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics 21 no.3*, 660-674.

Scholkopf, B., & Smola, A. J. (2001). *Learning with kernels: support vector machines, regularization, optimization, and beyond.* MIT press.

Statistics Solution. (2016). *Statistics Solution*. Retrieved August 14, 2019, from https://www.statisticssolutions.com/sample-size-calculation-and-sample-size-justification-resampling/

Suykens, J., & Vandewalle, J. (1999). Least Squares Support Vector Machine Classifiers. *Neural Processing Letters 9*.

Swets, J. A. (1988). Measuring the Accuracy of Diagnostic Systems. *Science, 240*(4857), 1285-1293.

Visa, S., & Ralescu, A. (2005, April). Issues in Mining Imbalanced Data Sets - A Review Paper. *Proceedings of the sixteen midwest artificial intelligence and cognitive science conference, 2005*, 67-73.

Wang, Y., Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change, 126*, 3-13.

Weglarz, G. (2004). Two Worlds of Data - Unstructured and Structured. *DM Review, 14*, 19-23.

Yen, S. J., & Lee, Y. S. (2006). Under-Sampling Approaches for Improving Prediction of the Minority Class in an Imbalanced Dataset. *Intelligent Control and Automation*, 731-740.

Yen, S. J., & Lee, Y. S. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications, 36*(3), 5718-5727.

Zhang, G. P. (2000). Neural Networks for Classification: A Survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 30*(4), 451-462.

Zhang, H. (2004). The Optimality of Naive Bayes. *American Association for Artificial Intelligence*.

Zhou, Z. H., & Liu, X. Y. (2006). Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem. *IEEE Transactions on Knowledge & Data Engineering*(1), 63-77.

Zurada, J. M. (1992). *Introduction to artificial neural systems.* St. Paul: West publishing company.

LAMPIRAN