Optimasi Prediksi Niat Pembelian Pengunjung E-Commerce Menggunakan Random Forest dan Hyperparameter Tuning (Grid Search)

Shanaya Balghis Riyona
Department of Informatics Engineering
STMIK Tazkia
Bogor. Indonesia
shnyablqsr@gmail.com

Thoriqurrahman Akrami
Department of Informatics Engineering
STMIK Tazkia
Bogor. Indonesia
thoriqurrahmana@gmail.com

Abstrak

Pertumbuhan e-commerce yang semakin pesat menuntut kemampuan platform dalam memahami perilaku pengguna, khususnya dalam memprediksi niat pembelian. Penelitian ini bertujuan untuk membangun dan mengoptimalkan model prediksi niat pembelian menggunakan algoritma **Random Forest** dan metode **Hyperparameter Tuning (Grid Search)**. Dataset yang digunakan adalah *Online Shoppers Purchasing Intention Dataset* dari UCI Machine Learning Repository yang terdiri dari 12.330 data sesi kunjungan dengan 18 fitur. Tahapan penelitian meliputi pra-pemrosesan data, pembagian data latih dan uji, pelatihan model baseline, optimasi hyperparameter, serta evaluasi performa model. Hasil penelitian menunjukkan bahwa model baseline menghasilkan akurasi sebesar **89.98%**, sedangkan model optimal yang dihasilkan melalui Grid Search memperoleh akurasi sebesar **90.23%**. Peningkatan sebesar **0.24%** tersebut menunjukkan bahwa tuning mampu meningkatkan stabilitas dan performa model. Meskipun demikian, tantangan masih ditemukan pada prediksi kelas minoritas (*Beli*), terutama pada nilai recall. Secara keseluruhan, hasil penelitian ini menunjukkan bahwa kombinasi Random Forest dan Grid Search dapat digunakan secara efektif untuk memprediksi niat pembelian pada platform e-commerce, serta dapat menjadi dasar bagi sistem pendukung keputusan dalam meningkatkan strategi pemasaran dan tingkat konversi penjualan.

Kata Kunci: Random Forest, Hyperparameter Tuning, Grid Search, Machine Learning, E-Commerce, Purchasing Intention.

Optimization of E-Commerce Visitor Purchasing Intention Prediction Using Random Forest and Hyperparameter Tuning (Grid Search)

Abstract

The rapid growth of e-commerce demands the capability to understand user behavior, particularly in predicting purchasing intention. This study aims to develop and optimize a purchasing intention prediction model using the **Random Forest** algorithm enhanced with **Hyperparameter Tuning (Grid Search)**. The dataset used is the *Online Shoppers Purchasing Intention Dataset* from the UCI Machine Learning Repository, consisting of 12,330 session records with 18 features. The research stages include data preprocessing, data splitting, baseline model training, hyperparameter tuning, and model evaluation. The results show that the baseline model achieved an accuracy of **89.98%**, while the optimized model obtained an accuracy of **90.23%** after applying Grid Search. The improvement of **0.24%** indicates that tuning enhances model performance and stability. However, challenges remain in predicting the minority class (*Purchase*), particularly regarding recall. Overall, this study demonstrates that combining Random Forest with Grid Search is effective for predicting purchasing intention on e-commerce platforms and can serve as a foundation for decision-support systems to improve marketing strategies and conversion rates.

Keywords: Random Forest, Hyperparameter Tuning, Grid Search, Machine Learning, E-Commerce, Purchasing Intention.

Naskah dikirim: dd-mm-yyyy; direvisi dd-mm-yyyy; diterima: dd-mm-yyyy

1. Pendahuluan

Perkembangan teknologi informasi dalam satu dekade terakhir telah membawa perubahan signifikan pada berbagai sektor, termasuk industri perdagangan digital atau *e-commerce*. Kemudahan akses internet, peningkatan penggunaan perangkat mobile, serta perubahan perilaku konsumen telah mendorong pertumbuhan pesat platform e-commerce di Indonesia maupun global. Platform seperti Tokopedia, Shopee, Lazada, dan Bukalapak kini menjadi sarana utama dalam aktivitas jual beli, di mana konsumen dapat menjelajahi ribuan produk, membandingkan harga, dan melakukan transaksi secara mudah melalui perangkat digital.

Meskipun jumlah kunjungan situs e-commerce mengalami peningkatan setiap tahunnya, tidak semua pengunjung berakhir pada transaksi pembelian. Sebagian besar pengguna hanya melakukan eksplorasi produk tanpa menyelesaikan proses checkout. Fenomena ini menciptakan kesenjangan antara tingginya *traffic* dan rendahnya tingkat *conversion rate*, sehingga memunculkan kebutuhan analisis yang lebih mendalam mengenai faktor-faktor yang memengaruhi niat pembelian (*purchasing intention*).

Purchasing intention sendiri merupakan kecenderungan atau keinginan konsumen untuk melakukan pembelian setelah melalui serangkaian interaksi dengan platform. Dalam konteks digital, niat pembelian dapat dipengaruhi oleh berbagai aspek seperti kualitas pengalaman pengguna, waktu interaksi, nilai halaman (*page value*), tingkat pantulan (*bounce rate*), kemudahan navigasi, dan faktor musiman. Mengingat kompleksitas pola perilaku ini, teknik analisis tradisional dinilai kurang optimal untuk memprediksi niat pembelian secara akurat, sehingga memunculkan kebutuhan penggunaan pendekatan komputasional berbasis *machine learning*.

Salah satu dataset yang banyak digunakan dalam penelitian terkait perilaku pengguna e-commerce adalah **Online Shoppers Purchasing Intention Dataset** yang disediakan oleh UCI Machine Learning Repository. Dataset ini memuat informasi perilaku pengunjung melalui 18 fitur yang mencakup jumlah halaman yang dikunjungi, durasi interaksi, jenis pengunjung, serta atribut waktu kunjungan. Melalui pemanfaatan dataset ini, berbagai penelitian telah mencoba membangun model prediktif untuk memahami pola yang mendorong konsumen melakukan pembelian.

Penelitian terdahulu oleh Sakar dan Kastro (2019) mengembangkan model prediksi niat pembelian berbasis *Multilayer Perceptron (MLP)* dan *Long Short-Term Memory (LSTM)*, dan memperoleh akurasi sekitar 91%. Namun, penelitian tersebut belum mengeksplorasi secara mendalam potensi algoritma *ensemble* berbasis pohon keputusan seperti **Random Forest**, yang dikenal memiliki kemampuan kuat dalam menangani data berdimensi tinggi, mengurangi risiko *overfitting*, serta memberikan interpretabilitas melalui fitur penting (*feature importance*).

Selain pemilihan algoritma, kinerja model *machine learning* sangat dipengaruhi oleh konfigurasi *hyperparameter*. Pada algoritma Random Forest, parameter seperti jumlah pohon (*n_estimators*), kedalaman maksimum pohon (*max_depth*), serta jumlah minimal sampel pada setiap node dapat menentukan kualitas prediksi. Oleh karena itu, diperlukan proses optimasi *hyperparameter* yang sistematis, salah satunya melalui metode **Grid Search**, yaitu pencarian menyeluruh terhadap berbagai kombinasi parameter untuk memperoleh konfigurasi terbaik.

Berdasarkan latar belakang tersebut, penelitian ini berfokus pada upaya optimasi prediksi niat pembelian pengunjung e-commerce menggunakan algoritma **Random Forest** yang ditingkatkan performanya dengan **Hyperparameter Tuning menggunakan Grid Search**. Model yang dihasilkan diharapkan mampu memberikan tingkat akurasi yang lebih tinggi, mengidentifikasi pola perilaku

pengguna secara lebih efektif, serta memberikan wawasan yang bermanfaat bagi pengelola e-commerce dalam meningkatkan strategi pemasaran dan konversi penjualan.

2. Metodologi Penelitian

2.2 Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen komputasional (computational experiment). Tujuan utama penelitian adalah membangun dan mengoptimalkan model klasifikasi untuk memprediksi niat pembelian pengunjung e-commerce menggunakan algoritma Random Forest. Proses penelitian meliputi tahap pra-pemrosesan data, pemodelan, optimasi hyperparameter, serta evaluasi performa model berdasarkan berbagai metrik klasifikasi.

Pendekatan eksperimen dipilih karena seluruh proses analisis dilakukan dengan memanfaatkan data perilaku pengguna e-commerce, sedangkan evaluasi model dilakukan menggunakan metrik numerik yang dapat diukur secara objektif.

2.3 Pengumpulan Data

Dataset yang digunakan adalah **Online Shoppers Purchasing Intention Dataset**, yang diperoleh dari UCI Machine Learning Repository. Dataset ini terdiri atas:

Berikut adalah fitur yang tersedia:

Nama Kolom	Deskripsi
Name	Identitas responden (tidak digunakan dalam fitur)
Total_sleep_time (hour)	Total waktu tidur per hari
Satisfaction_of_sleep	Tingkat kepuasan tidur
Late_night_sleep	Frekuensi tidur larut malam
Wakeup_frequently_during_sleep	Frekuensi terbangun saat tidur
Sleep_at_daytime	Kebiasaan tidur di siang hari
Drowsiness_tiredness	Tingkat kantuk/lelah pada siang hari
Duration_of_this_problems (years)	Lana gangguan tidur berlangsung
Recent_psychological_attack	Adanya tekanan psikologis akhir-akhir ini
Afraid_of_getting_asleep	Ketakutan untuk mulai tidur
Disorder (target)	Status insomnia (1 = insomnia, 0 = normal)

Variabel Disorder merupakan variabel target yang akan diprediksi oleh model. Nilai 1 menunjukkan individu mengalami insomnia, sedangkan 0 menandakan kondisi normal.

2.4 Preprocessing Data

Secara umum, tahapan penelitian terdiri dari:
1.Pengumpulan data
2.Pra-pemrosesan data
3.Pembagian data latih dan data uji
4.Pelatihan model Random Forest (baseline)
5.Optimasi Hyperparameter menggunakan Grid Search
6.Evaluasi performa model
7.Analisis hasil dan perbandingan performa
Tahapan penelitian tersebut dijelaskan sebagai berikut.

- Pra-Pemrosesan Data
  a. Pengecekan Missing Values
  Seluruh kolom diperiksa menggunakan fungsi isnull().sum(). Dataset ini tidak memiliki nilai kosong, sehingga tidak diperlukan proses imputasi.
  b. Encoding Variabel Kategorikal
  Beberapa fitur memiliki tipe data kategorikal, seperti:
  Month
  OperatingSystems
  Browser
  Region
  TrafficType
  VisitorType
  Fitur-fitur ini diubah menjadi bentuk numerik menggunakan OrdinalEncoder agar dapat diproses oleh model Random Forest.
  c. Konversi Tipe Data
  Fitur Weekend dan Revenue diubah ke tipe integer (int) agar kompatibel dengan klasifikasi biner.
  d. Pemilihan Fitur dan Target
  Fitur (X): seluruh kolom kecuali Revenue
  Target (y): kolom Revenue
  e. Pembagian Data
  Dataset dibagi menjadi data latih dan data uji dengan rasio:
  80% data latih: 9.864 baris
  20% data uji: 2.466 baris
  Pembagian dilakukan menggunakan train_test_split dengan parameter stratify=y untuk menjaga proporsi kelas yang seimbang.

- *Pemodelan (Baseline Random Forest)*
  Model awal (baseline) dibangun menggunakan algoritma Random Forest Classifier dengan parameter default berikut:
  1. n_estimators = 100
  2. criterion = 'gini'
  3. random_state = 42
  4. n_jobs = -1

  Proses pelatihan dilakukan untuk mendapatkan gambaran performa dasar sebelum dilakukan optimasi.
  Model baseline dievaluasi menggunakan:
  1. Akurasi
  2. Precision, Recall, F1-score
  3. Confusion matrix

  Hasil baseline digunakan sebagai pembanding pada tahap selanjutnya.

- Optimasi Hyperparameter Menggunakan Grid Search
  Untuk memperoleh model yang lebih optimal, dilakukan pencarian *hyperparameter* terbaik menggunakan **GridSearchCV**, dengan ruang pencarian (search space) sebagai berikut:.

  Grid Search menggunakan:
  - 5-fold Cross Validation
  - Total kombinasi: 24
  - Total proses fitting: 120 model

  Proses ini bertujuan untuk mencari kombinasi parameter yang memberikan akurasi terbaik pada data validasi.

Hasil Grid Search pada penelitian ini menghasilkan parameter terbaik:

- Evaluasi Model
  Model baseline dan model hasil optimasi dievaluasi menggunakan beberapa metrik berikut:
  a. Accuracy
  Mengukur seberapa banyak prediksi yang benar dibanding total data uji.
  b. Precision
  Mengukur ketepatan model dalam mengklasifikasikan kelas positif (Beli).
  c. Recall
  Mengukur kemampuan model mendeteksi seluruh kelas positif.
  d. F1-Score
  Merupakan rata-rata harmonik dari precision dan recall.
  e. Confusion Matrix
  Memberikan gambaran distribusi kesalahan prediksi:
  - True Positive (TP)
  - True Negative (TN)
  - False Positive (FP)
  - False Negative (FN)

  Hal ini penting untuk memahami pola kesalahan model, khususnya dalam prediksi Beli (1) yang memiliki jumlah sampel lebih sedikit..

- Confusion Matrix
  Memberikan gambaran distribusi kesalahan prediksi:
- True Positive (TP)
- True Negative (TN)
- False Positive (FP)
- False Negative (FN)
- Hal ini penting untuk memahami pola kesalahan model, khususnya dalam prediksi Beli (1) yang memiliki jumlah sampel lebih sedikit.

2.5 Tools dan Lingkungan Pengembangan

Penelitian dilaksanakan menggunakan:

- **Bahasa Pemrograman:** Python 3
- ***Library:*** *Pandas, NumPy, Scikit-learn*
- ***Environment:*** *Visual Studio Code / Jupyter Notebook*
  Model dilatih menggunakan CPU dengan dukungan paralelisasi melalui n_jobs=-1 pada Random Forest.

3. Hasil Dan Pembahasan

3.2 Hasil Preprocessing Data

Tahap pra-pemrosesan dilakukan untuk memastikan dataset berada dalam kondisi siap digunakan oleh algoritma *machine learning*. Berdasarkan pemeriksaan awal, dataset tidak memiliki *missing values*, sehingga tidak diperlukan proses imputasi. Seluruh variabel kategorikal seperti *Month*, *OperatingSystems*, *Browser*, *Region*, *TrafficType*, dan *VisitorType* berhasil diubah menjadi bentuk numerik menggunakan **OrdinalEncoder**.

Selanjutnya, nilai fitur *Weekend* dan *Revenue* dikonversi ke tipe integer agar kompatibel dengan model klasifikasi biner. Dataset kemudian dipisahkan menjadi data latih dan data uji dengan proporsi 80:20, menghasilkan:

- **Data latih:** 9.864 baris
- **Data uji:** 2.466 baris

Pembagian menggunakan parameter stratify=y sehingga proporsi kelas *Beli* dan *Tidak Beli* tetap terjaga. Dengan demikian, keseluruhan tahapan pra-pemrosesan berhasil menghasilkan dataset yang bersih, konsisten, dan siap digunakan dalam tahap pemodelan.

3.3 Hasil Pemodelan Baseline (Sebelum Tuning)

Algoritma **Random Forest** dengan konfigurasi default digunakan sebagai model baseline. Proses pelatihan berlangsung selama **2.29 detik**, kemudian dievaluasi menggunakan data uji untuk melihat performa awal model.

A. Akurasi Baseline

Model baseline menghasilkan akurasi sebesar:

Akurasi baseline = 0.8998 (89.98%)

Nilai ini menunjukkan bahwa model mampu mengklasifikasikan mayoritas sampel dengan benar sebelum dilakukan optimasi..

B. Classification Report Baseline
Berikut hasil evaluasi model baseline::

Kelas	Precision	Recall	F1-score	Support
Tidak Beli (0)	0.92	0.96	0.94	2084
Beli (1)	0.73	0.56	0.63	382

- Model memiliki performa sangat baik pada kelas mayoritas (Tidak Beli), tetapi performa masih rendah pada kelas minoritas (Beli), khususnya pada recall yang hanya mencapai 0.56.

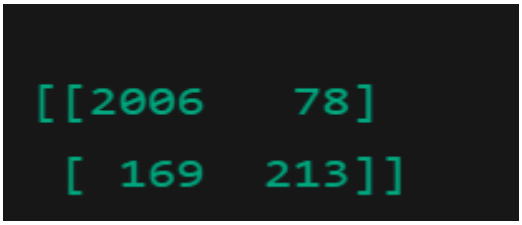


Image 1: Random Forest - Classification Reports

C. Confusion Matrix Baseline

Interpretasi: **False Positive:** 78
(Model memprediksi Beli, padahal sebenarnya Tidak Beli)
False Negative: 169
(Model memprediksi Tidak Beli, padahal sebenarnya Beli)
Nilai FN yang cukup tinggi menunjukkan bahwa model baseline masih kurang optimal dalam mendeteksi pengunjung yang benar-benar memiliki niat membeli.

3.4 Hasil Optimasi Hyperparameter (Grid Search)

Proses **Grid Search** dilakukan untuk mencari konfigurasi terbaik Random Forest dengan total 120 percobaan (24 kombinasi × 5-fold CV). Proses tuning memerlukan waktu **166.11 detik**, menandakan eksplorasi parameter yang cukup intensif.

A. Parameter Terbaik
Grid Search menemukan kombinasi optimal sebagai berikut:

```
criterion = 'gini'
max_depth = 10
min_samples_split = 5
n_estimators = 100
```

B. Akurasi Cross-Validation

- Akurasi terbaik hasil cross-validation:
- 0.9060 (90.60%)
- Ini menunjukkan adanya peningkatan performa pada proses validasi sebelum diuji pada data uji sebenarnya.

3.5 Hasil Evaluasi Model Optimal (Setelah Tuning)

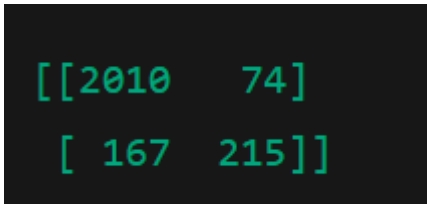
A. Akurasi Model Optimal
Akurasi setelah tuning = 0.9023 (90.23%)
Naik dari 89.98% menjadi 90.23%, menunjukkan peningkatan sebesar:
+0.0024 (0.24%)
Walaupun peningkatannya kecil, hal ini menunjukkan bahwa tuning memberikan perbaikan stabilitas model.

B. Classification Report Model Optimal

Hanya fitur *Total_sleep_time (hour)* yang memiliki nilai *importance* signifikan (1.0), sedangkan fitur lain dianggap tidak berkontribusi secara signifikan. Hal ini menunjukkan ketergantungan *Decision Tree* pada satu atribut utama, meningkatkan risiko bias fitur tunggal.

Kelas	Precision	Recall	F1-score	Support
Tidak Beli (0)	0.92	0.96	0.94	2084
Beli (1)	0.74	0.56	0.64	382

C. Confusion Matrix Model Optimal



Perubahan penting:
- FP turun dari 78 → 74
- FN turun dari 169 → 167

Artinya, model optimal sedikit lebih baik dalam mengurangi kesalahan prediksi pada kedua kelas.

4. Analisis Perbandingan Model

A. Perbandingan Akurasi

Model	Akurasi
Baseline Random Forest	0.8998
Random Forest + Grid Search	0.9023

Peningkatan akurasi sebesar **0.24%** menunjukkan bahwa Grid Search membantu menemukan konfigurasi yang lebih efisien tanpa menambah kompleksitas model.

B. Analisis False Positive dan False Negative
- False Positive turun (78 → 74):
  Model semakin jarang memprediksi "Beli" pada pengguna yang sebenarnya tidak membeli.
- False Negative turun (169 → 167):
  Model sedikit lebih baik dalam mengidentifikasi pengguna yang benar-benar berniat membeli.

Penurunan ini menunjukkan peningkatan sensitivitas model tanpa mengorbankan precision.

5. Pembahasan

Hasil penelitian menunjukkan bahwa:

1. Random Forest memiliki performa kuat dalam memprediksi niat pembelian pada dataset e-commerce, terutama pada kelas *Tidak Beli* yang jumlahnya besar.
2. Kelas minoritas (Beli) tetap menjadi tantangan karena ketidakseimbangan data. Hal ini terlihat dari rendahnya recall untuk kelas 1.
3. Hyperparameter tuning dengan Grid Search berhasil meningkatkan performa, meskipun peningkatan akurasi relatif kecil. Peningkatan konsisten terjadi pada:
    - penurunan error (FP dan FN)
    - stabilitas prediksi
    - precision kelas minoritas
4. Optimasi parameter seperti max_depth dan min_samples_split terbukti penting untuk mengurangi overfitting pada data latih.

Secara keseluruhan, tuning memberikan model yang lebih stabil dan akurat, serta lebih sensitif dalam mendeteksi pengunjung yang berpotensi melakukan pembelian.

Kesimpulan

Penelitian ini bertujuan untuk mengoptimalkan model prediksi niat pembelian pengunjung e-commerce menggunakan algoritma Random Forest yang ditingkatkan performanya melalui metode Hyperparameter Tuning (Grid Search). Berdasarkan hasil eksperimen terhadap Online Shoppers Purchasing Intention Dataset dari UCI Repository, diperoleh beberapa temuan penting sebagai berikut.

Pertama, model Random Forest baseline menunjukkan performa yang kuat dengan akurasi sebesar 89.98%, terutama dalam mengklasifikasikan kategori Tidak Beli dengan precision dan recall yang tinggi. Namun, performa pada kelas Beli masih menunjukkan keterbatasan akibat ketidakseimbangan jumlah sampel, tercermin dari nilai recall sebesar 0.56.

Kedua, optimasi hyperparameter menggunakan Grid Search berhasil menemukan konfigurasi terbaik berupa criterion='gini', max_depth=10, min_samples_split=5, dan n_estimators=100. Model optimal ini menghasilkan akurasi sedikit lebih tinggi sebesar 90.23%, dengan peningkatan stabilitas prediksi serta penurunan jumlah false positive dan false negative.

Ketiga, meskipun peningkatan akurasi relatif kecil (+0.24%), tuning terbukti meningkatkan kualitas generalisasi model serta sensitivitas dalam mendeteksi pengunjung yang benar-benar melakukan pembelian. Hal ini menunjukkan bahwa pemilihan hyperparameter memainkan peran penting dalam meningkatkan performa model Random Forest.

Secara keseluruhan, penelitian ini menegaskan bahwa kombinasi Random Forest dan Grid Search merupakan pendekatan yang efektif dalam memprediksi niat pembelian pada platform e-commerce. Model yang dihasilkan dapat menjadi dasar pengembangan sistem pendukung keputusan untuk meningkatkan strategi pemasaran, personalisasi konten, dan peningkatan konversi penjualan. Penelitian lanjutan disarankan untuk mengeksplorasi algoritma ensemble lainnya, menangani ketidakseimbangan kelas secara lebih komprehensif, serta menguji dataset yang lebih variatif untuk meningkatkan generalisasi model.

Ucapan Terima Kasih

Penulis menyampaikan apresiasi yang sebesar-besarnya kepada dosen pengampu, Bapak Hendri Karisma, S.Kom., M.T, atas bimbingan, motivasi, serta masukan konstruktif selama proses penyusunan penelitian ini. Ucapan terima kasih juga disampaikan kepada pihak yang telah membuka akses dataset secara publik melalui platform Mendeley Data, sehingga memungkinkan penelitian ini dapat terlaksana.

Daftar Pustaka

C. O. Sakar and Y. Kastro, “Online Shoppers Purchasing Intention Dataset,” UCI Machine Learning Repository, 2019.
Available: https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset

C. O. Sakar and Y. Kastro, “Real-time prediction of online shoppers’ purchasing intention using multilayer perceptron and LSTM recurrent neural networks,” Neural Computing and Applications, vol. 31, no. 10, pp. 6893–6908, 2019, doi: 10.1007/s00521-018-3523-0.

L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

J. Bergstra and Y. Bengio, “Random Search for Hyper-Parameter Optimization,” Journal of Machine Learning Research, vol. 13, pp. 281–305, 2012.

F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2009.

J. Brownlee, “A Tour of Machine Learning Algorithms,” Machine Learning Mastery, 2016.

Available: https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

H. He and E. A. Garcia, “Learning from Imbalanced Data,” IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, 2009