KOLANO: Journal of Multi-Disciplinary Sciences
Vol 1, Issue 1, September 2020, Page. 1-6
ISSN: 2723-4312 (Media Online)
Available at: https://ejournal.sangadjimediapublishing.id/index.php/kolano

Optimasi Penilaian Kredit Nasabah dengan Metode Expectation-Maximization-Naïve Bayes

Raditia Vindua

Program Studi Teknik Informatika, Fakultas Teknik Informatika, Universitas Pamulang, Indonesia

Email: dosen02380@unpam.ac.id

Abstract. Customer credit data has not been optimally utilized to identify patterns that can be used to predict the eligibility of new credit applicants. One of the main challenges is the absence of class labels in customer credit data, which hinders the classification process. This study aims to develop a data mining model that combines the EM (Expectation-Maximization) clustering method and Naïve Bayes classification to predict the eligibility of new credit applicants. The EM clustering method is used to assign class labels to unlabeled data, enabling the classification process with Naïve Bayes. From a total of 540 customer credit data points analyzed, 142 were classified into cluster 0 (non-eligible credit applicants) and 398 into cluster 1 (eligible credit applicants). The results indicate that the combined method achieved an average accuracy rate of 99.24% and an average error rate of 0.76%. Using the WEKA software, this study concludes that the combination of EM clustering and Naïve Bayes classification is an effective approach for predicting the eligibility of new credit applicants.

Keywords: Credit Client Eligibility, Data Mining, EM Clustering, Naïve Bayes Classification.

Abstrak. Data nasabah kredit belum dimanfaatkan secara optimal untuk mengenali pola yang dapat digunakan dalam memprediksi kelayakan calon nasabah kredit baru. Salah satu tantangan utama adalah tidak adanya label kelas pada data nasabah kredit yang memungkinkan proses klasifikasi. Penelitian ini bertujuan untuk mengembangkan model data mining yang menggabungkan metode klasterisasi EM (Expectation-Maximization) dan klasifikasi Naïve Bayes guna memprediksi kelayakan nasabah kredit baru. Klasterisasi EM digunakan untuk memberikan label kelas pada data yang tidak berlabel, sehingga memungkinkan proses klasifikasi dengan Naïve Bayes. Dari total 540 data nasabah kredit yang dianalisis, sebanyak 142 data tergolong dalam cluster 0 (nasabah kredit tidak layak) dan 398 data dalam cluster 1 (nasabah kredit layak). Hasil pengujian menunjukkan bahwa metode gabungan ini menghasilkan tingkat akurasi rata-rata sebesar 99,24% dan rata-rata tingkat error sebesar 0,76%. Dengan menggunakan perangkat lunak WEKA, penelitian ini menyimpulkan bahwa kombinasi klasterisasi EM dan klasifikasi Naïve Bayes merupakan pendekatan yang efektif dalam memprediksi kelayakan nasabah kredit baru.

Kata kunci: Kelayakan Nasabah Kredit, Data Mining, Klasterisasi EM, Klasifikasi Naïve Bayes

1. Pendahuluan

Permasalahan dalam pemberian kredit pinjaman sering kali terjadi pada perusahaan keuangan. Salah satu isu yang menjadi penyebab adalah nasabah yang mengalami keterlambatan dalam membayar angsuran. Pada awalnya, nasabah tersebut dianggap layak menerima kredit, tetapi seiring waktu, mereka berpotensi tinggi untuk mengalami keterlambatan pembayaran, misalnya karena meningkatnya tanggungan finansial atau kewajiban angsuran lainnya. Untuk mengurangi risiko kredit macet, diperlukan analisis data nasabah yang sebelumnya telah dikumpulkan. Data ini dikenal sebagai data pelatihan (training data), yang dapat diubah menjadi decision

tree atau pola lain yang mudah dimengerti sehingga membantu dalam menentukan apakah pengajuan kredit diterima atau ditolak.

Berdasarkan permasalahan ini, penelitian ini bertujuan untuk memanfaatkan data nasabah kredit untuk mengenali pola (pattern) yang dapat memprediksi kelayakan calon nasabah kredit baru. Namun, data nasabah kredit yang tersedia belum memiliki label kelas untuk diklasifikasikan. Oleh karena itu, penelitian ini menggunakan algoritma Expectation Maximization (EM) untuk memberi label kelas sebagai langkah awal untuk membantu algoritma Naïve Bayes dalam pengklasifikasian. Naïve Bayes dikenal sebagai algoritma yang sederhana dan efisien. Dalam penelitian Astrid, algoritma ini diterapkan pada data siswa untuk memprediksi prestasi. Dengan menggunakan 21 atribut dalam proses prediksi, hasil akurasi mencapai 95,69%, lebih tinggi dibandingkan algoritma C4.5 yang hanya mencapai akurasi 90,95% [1].

Penelitian ini menggunakan data nasabah kredit dari MNC Finance cabang Alam Sutera pada tahun 2016. Data diolah menggunakan perangkat lunak Microsoft Excel dan WEKA, sebuah perangkat lunak untuk machine learning. WEKA menyediakan berbagai fungsi untuk klasifikasi dan klasterisasi data yang dapat disesuaikan dengan kebutuhan pengguna. Tujuan utama penelitian ini adalah untuk menerapkan algoritma Naïve Bayes dalam memprediksi kelayakan pemberian kredit kepada nasabah. Hasil yang diharapkan adalah tingkat akurasi yang tinggi dalam menentukan kelayakan pemberian kredit.

2. Tinjauan Pustaka

Penilaian kredit merupakan elemen penting dalam sistem keuangan modern karena membantu lembaga keuangan dalam menilai risiko finansial dan membuat keputusan strategis terkait pemberian kredit. Sistem penilaian kredit memanfaatkan data historis untuk memprediksi kemungkinan calon nasabah gagal membayar. Probabilitas yang diperoleh dibandingkan dengan nilai ambang batas tertentu, di mana kredit diberikan jika nilai prediksi berada di bawah ambang tersebut dan ditolak jika sebaliknya [5]. Dalam dekade terakhir, perkembangan teknologi data telah memungkinkan adopsi teknik data mining untuk mendukung proses penilaian kredit.

Data mining didefinisikan sebagai proses penggalian informasi penting yang tersembunyi dalam data besar yang tidak terstruktur. Teknik ini melibatkan tahapan identifikasi masalah bisnis, pengambilan data yang relevan, dan analisis data untuk menemukan pola yang signifikan bagi pengambilan keputusan strategis [2]. Dalam konteks penilaian kredit, teknik data mining seperti klasifikasi dan klasterisasi sering digunakan untuk menganalisis data nasabah dan memprediksi kelayakan kredit. Klasifikasi, misalnya, bertujuan untuk membangun model yang dapat memprediksi hubungan antara variabel independen dan label kelas tertentu, seperti "akun baik" atau "akun buruk" [3].

Naïve Bayes adalah salah satu algoritma klasifikasi yang populer karena kesederhanaannya dan efisiensinya dalam menangani dataset besar. Algoritma ini didasarkan pada teorema Bayes dengan asumsi independensi antar variabel. Han dan Kamber [4] menunjukkan bahwa Naïve Bayes tidak hanya cepat dalam proses pelatihan data tetapi juga memiliki akurasi tinggi pada dataset besar. Penelitian Astrid [1] menunjukkan bahwa algoritma Naïve Bayes mampu mencapai akurasi hingga 95,69% dalam memprediksi prestasi siswa berdasarkan 21 atribut, yang lebih tinggi dibandingkan dengan algoritma C4.5 yang hanya mencapai akurasi 90,95%. Efektivitas Naïve Bayes menjadikannya pilihan utama dalam berbagai aplikasi, termasuk prediksi kelayakan kredit.

Di sisi lain, klasterisasi digunakan untuk mengelompokkan data tanpa label ke dalam kelompok homogen berdasarkan kemiripan tertentu. Metode ini bertujuan untuk meminimalkan perbedaan antar data dalam satu kelompok dan memaksimalkan perbedaan antar kelompok [6]. Klasterisasi sering digunakan sebagai langkah awal dalam proses klasifikasi, terutama ketika data yang tersedia tidak memiliki label. Algoritma Expectation Maximization (EM) merupakan salah satu metode klasterisasi yang efektif karena kemampuannya dalam menangani data dengan missing values dan menghasilkan hasil yang akurat. Osama [7] menunjukkan bahwa algoritma EM mampu mencapai tingkat akurasi 90% pada dataset besar, lebih tinggi dibandingkan K-Means yang hanya mencapai 89%.

Pendekatan semi-supervised learning telah menjadi solusi inovatif dalam memanfaatkan data tanpa label. Pendekatan ini menggabungkan klasterisasi dan klasifikasi untuk memaksimalkan informasi yang tersedia. Misalnya, data tanpa label dikelompokkan menggunakan algoritma EM, dan hasilnya digunakan sebagai input untuk proses klasifikasi menggunakan Naïve Bayes. Kombinasi teknik ini terbukti meningkatkan akurasi prediksi dibandingkan dengan metode konvensional [8]. Selain itu, beberapa penelitian telah mengembangkan pendekatan hibrida untuk meningkatkan akurasi penilaian kredit. Lee et al. [9] mengusulkan kombinasi BP Neural Network dan analisis diskriminan yang menunjukkan tingkat konvergensi dan akurasi yang lebih tinggi. Huang et al. [10] mengintegrasikan Support Vector Machines (SVM), algoritma genetika, dan F-score untuk meningkatkan efisiensi prediksi. Penelitian ini menekankan pentingnya eksplorasi metode hibrida dalam mengatasi kompleksitas data kredit.

Dalam konteks implementasi, perangkat lunak seperti WEKA telah banyak digunakan untuk analisis data. WEKA menyediakan berbagai algoritma untuk klasterisasi, klasifikasi, dan prediksi yang dapat disesuaikan dengan kebutuhan pengguna. Han dan Kamber [4] menyebutkan bahwa WEKA adalah alat yang fleksibel dan mudah diimplementasikan dalam berbagai studi kasus, termasuk penilaian kredit. Penelitian ini menggunakan WEKA untuk mengolah data nasabah kredit dan menerapkan algoritma EM serta Naïve Bayes guna memprediksi kelayakan kredit dengan akurasi tinggi. Clustering memainkan peran penting dalam mengidentifikasi pola tersembunyi dalam data. Teknik ini memungkinkan sistem prediksi yang lebih transparan dan dapat dipahami. Xiong et al. [11] menjelaskan bahwa data clustering dapat digunakan untuk membangun pola yang berfungsi sebagai dasar dalam sistem prediksi akhir. Teknik ini tidak hanya meningkatkan interpretabilitas tetapi juga memperbaiki hasil prediksi.

Algoritma EM memiliki keunggulan dalam mengelompokkan data secara optimal, bahkan dalam kondisi data yang kompleks. Ian dan Iebe [12] menunjukkan bahwa algoritma EM dapat menemukan parameter model dengan probabilitas tinggi pada setiap iterasi, menjadikannya pilihan yang andal dalam analisis data. Selain itu, Acock [6] menyatakan bahwa EM mampu menghasilkan klasterisasi yang lebih baik dibandingkan metode lainnya, terutama pada dataset dengan skala besar. Dengan memanfaatkan kombinasi teknik EM untuk klasterisasi dan Naïve Bayes untuk klasifikasi, penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam meningkatkan akurasi prediksi kelayakan kredit. Pendekatan ini tidak hanya efektif dalam mengolah data tetapi juga efisien dalam mendukung pengambilan keputusan strategis. Penggunaan data historis yang optimal melalui teknik data mining membuka peluang baru untuk meningkatkan kualitas penilaian kredit dan meminimalkan risiko finansial bagi lembaga keuangan.

3. Metode

Penelitian ini menggunakan pendekatan Cross Industry Standard Process for Data Mining (CRISP-DM), yang terdiri dari enam tahapan utama. Tahap pertama adalah Business Understanding, yang bertujuan untuk memahami dan menganalisis proses bisnis dari sistem yang diteliti. Pada fase ini, peneliti menentukan tujuan, yaitu prediksi kelayakan nasabah kredit, serta merancang strategi untuk mencapai tujuan tersebut [13]. Selanjutnya, pada tahap Data Understanding, dilakukan proses pembelajaran terhadap data dengan mengumpulkan informasi dari berbagai sumber, seperti buku panduan dan internet, untuk memahami metode yang relevan, termasuk metode klasterisasi Expectation Maximization (EM) dan metode klasifikasi Naïve Bayes [14].

Tahap berikutnya adalah Data Preparation, yang melibatkan seleksi, pembersihan, integrasi, dan transformasi data agar siap digunakan dalam pemodelan [15]. Setelah data siap, proses dilanjutkan ke tahap Modeling, di mana dataset dimodelkan menggunakan algoritma klasterisasi. Model klasterisasi yang dihasilkan kemudian dianalisis lebih lanjut dengan algoritma klasifikasi Naïve Bayes, sehingga menghasilkan parameter evaluasi beserta nilai akurasi [16]. Pada tahap Evaluation, dilakukan analisis kuantitatif untuk mengevaluasi hasil pemodelan dengan membandingkan akurasi hasil klasifikasi guna memastikan kualitas model yang dikembangkan [17]. Tahapan terakhir adalah Deployment, yang berfokus pada penyusunan laporan atau presentasi untuk menyampaikan wawasan dan pengetahuan yang diperoleh dari proses evaluasi [18]. Keseluruhan tahapan ini memastikan bahwa proses data mining dilakukan secara terstruktur dan menghasilkan hasil yang relevan dengan tujuan penelitian.

4. Hasil dan Pembahasan

Penelitian ini menggunakan data Nasabah Kredit MNC Finance cabang Alam Sutera pada tahun 2016, dengan total 540 data nasabah. Data ini kemudian diolah menggunakan Microsoft Excel dengan atribut sebagai berikut: Kota, Umur, Status Pernikahan, Lama Bekerja, Jenis Pekerjaan, Jumlah Tanggungan, Status Rumah, Overdue (pembayaran yang melewati batas waktu), Overdue Day (OD), dan Denda. Selanjutnya, data disimpan dalam format *.csv agar dapat diproses menggunakan aplikasi WEKA. Metodologi pemodelan data mining yang digunakan adalah kombinasi antara klasterisasi Expectation Maximization (EM) dan klasifikasi Naïve Bayes.

4.1 Klasterisasi EM

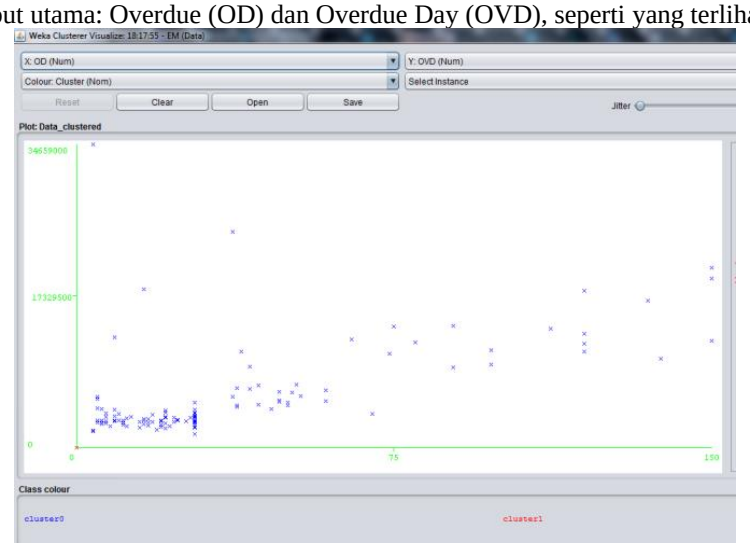
Tahap awal pemrosesan data dilakukan melalui metode klasterisasi EM. Metode ini memanfaatkan estimasi Maximum Likelihood (ML) untuk mengelompokkan data berdasarkan parameter probabilistik. Hasil dari proses klasterisasi ini divisualisasikan pada Gambar 1 yang menunjukkan log likelihood sebesar -41,28322. Berdasarkan hasil klasterisasi, terdapat dua klaster utama:

a) Cluster 0: Berisi 142 data nasabah.
b) Cluster 1: Berisi 398 data nasabah.

```
Time taken to build model (full training data) : 0.21 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       142 ( 26%)
1       398 ( 74%)


Log likelihood: -41.28322
```

Gambar 1 Hasil *Clustering EM* menggunakan *WEKA*

Setiap data dikelompokkan ke dalam salah satu klaster berdasarkan karakteristiknya. Untuk menentukan kualitas masing-masing klaster, analisis lebih lanjut dilakukan dengan memvisualisasikan posisi cluster berdasarkan dua atribut utama: Overdue (OD) dan Overdue Day (OVD), seperti yang terlihat pada Gambar 2.



Gambar 2 Visualisasi *Cluster EM* berdasarkan *OD* dan *OVD*

Berdasarkan visualisasi tersebut:

a) Cluster 1 (warna merah): Berlokasi di koordinat (0,0), menunjukkan nasabah yang tidak memiliki tunggakan pembayaran maupun keterlambatan hari. Hal ini mengindikasikan bahwa Cluster 1 adalah kelompok nasabah yang baik.
b) Cluster 0 (warna biru): Tersebar di area di mana x > 0 dan y > 0, yang artinya nasabah dalam klaster ini memiliki nilai overdue dan overdue day yang lebih besar dari nol. Dengan demikian, Cluster 0 merepresentasikan nasabah yang memiliki risiko kredit buruk.

Kesimpulan dari analisis ini adalah bahwa nasabah dalam Cluster 1 dianggap sebagai nasabah yang layak atau memenuhi kriteria baik, sedangkan nasabah dalam Cluster 0 merupakan nasabah yang berisiko tinggi.

4.2 Klasifikasi Naïve Bayes

Setelah klasterisasi dilakukan, tahap berikutnya adalah menerapkan algoritma klasifikasi Naïve Bayes. Algoritma ini digunakan untuk memprediksi kategori nasabah berdasarkan pola yang telah ditemukan pada tahap klasterisasi. Hasil klasifikasi Naïve Bayes dibandingkan dengan hasil klasterisasi EM untuk mengukur tingkat akurasi. Detail hasil perbandingan ditampilkan pada Tabel 1.

Tabel 1. Hasil Perbandingan Klasifikasi *Naïve Bayes* dengan Klasterisasi *EM*

Test Options Naïve Bayes	***Naïve Bayes* dengan *EM***	
	Akurasi	***Error***
Use Training Set	99.4444%	0.5556%
Supplied Test Set	98.0861%	1.9139%
Cross-validation 10-fold	99.4444%	0.5556%
Percentage 75%	100%	0.0000%

Berdasarkan tabel tersebut, tingkat akurasi rata-rata dari kombinasi Naïve Bayes dengan klasterisasi EM mencapai 99.2437%, dengan tingkat error rata-rata sebesar 0.7563%. Hasil ini menunjukkan bahwa kombinasi metode tersebut memberikan prediksi yang sangat akurat untuk menentukan kelayakan nasabah.

4.3 Evaluasi

Evaluasi hasil klasterisasi dan klasifikasi menunjukkan bahwa kombinasi metode EM dan Naïve Bayes mampu menghasilkan model yang andal dalam mengelompokkan nasabah berdasarkan kualitas kredit. Tingginya akurasi model ini memperkuat validitas pendekatan yang digunakan, terutama dalam konteks pengelolaan risiko kredit.

Untuk menguji keandalan model, penelitian ini menggunakan 20 data nasabah kredit baru yang belum terklasifikasi. Data ini digunakan sebagai studi kasus untuk memprediksi kelayakan nasabah mendapatkan pinjaman kredit. Berdasarkan hasil prediksi, sebanyak 13 nasabah diterima sementara 7 nasabah ditolak. Pola yang ditemukan dari analisis data menunjukkan bahwa:

a) Nasabah dengan jenis pekerjaan wiraswasta (entrepreneur) yang memiliki lama bekerja kurang dari 10 tahun cenderung ditolak untuk mendapatkan pinjaman kredit.
b) Sebaliknya, nasabah dengan pekerjaan yang lebih stabil atau lama bekerja lebih dari 10 tahun memiliki peluang lebih besar untuk diterima.

Hasil ini sejalan dengan temuan sebelumnya yang menunjukkan bahwa karakteristik pekerjaan dan pengalaman kerja memiliki pengaruh signifikan terhadap kualitas kredit nasabah.

4.4 Deployment

Tahap terakhir dalam proses ini adalah deployment, yaitu penyusunan laporan hasil analisis dan rekomendasi berdasarkan pola yang ditemukan. Pengetahuan yang diperoleh dari proses data mining ini dapat dimanfaatkan untuk mendukung pengambilan keputusan strategis oleh perusahaan, khususnya dalam mengelola risiko kredit. Beberapa langkah yang dapat dilakukan oleh MNC Finance berdasarkan hasil penelitian ini adalah:

a) Mengembangkan Sistem Evaluasi Kredit: Perusahaan dapat mengintegrasikan model klasterisasi EM dan Naïve Bayes ke dalam sistem evaluasi kredit untuk memberikan rekomendasi otomatis terkait kelayakan nasabah.
b) Meningkatkan Pengawasan pada Nasabah Berisiko Tinggi: Nasabah yang termasuk dalam Cluster 0 perlu mendapatkan perhatian khusus, misalnya dengan menawarkan program restrukturisasi atau monitoring intensif.
c) Menyusun Kebijakan Kredit yang Lebih Selektif: Berdasarkan pola yang ditemukan, perusahaan dapat menyusun kebijakan kredit yang lebih selektif untuk nasabah dengan pekerjaan wiraswasta dan lama bekerja kurang dari 10 tahun.

5. Kesimpulan

Berdasarkan analisis yang dilakukan, dapat disimpulkan bahwa metode gabungan klasterisasi EM dan klasifikasi Naïve Bayes menggunakan WEKA efektif dalam memprediksi kelayakan nasabah kredit. Dari 540 data nasabah, klasterisasi EM menghasilkan dua kelompok, yaitu cluster 1 (398 nasabah kredit baik) dan cluster 0 (142 nasabah kredit buruk). Metode ini menunjukkan tingkat akurasi klasifikasi yang tinggi, dengan rata-rata akurasi 99,2437% dan rata-rata error hanya 0,7563%. Selain itu, hasil prediksi kelayakan terhadap 20 data nasabah kredit baru menunjukkan bahwa 13 nasabah diterima dan 7 nasabah ditolak. Pola menarik ditemukan, di mana nasabah dengan jenis pekerjaan wiraswasta dan masa kerja di bawah 10 tahun cenderung ditolak. Dengan demikian, metode ini terbukti mampu memberikan hasil yang andal dalam mendukung keputusan terkait kelayakan kredit nasabah.

Referensi

[1] Astrid, “Penerapan Naïve Bayes untuk prediksi prestasi siswa,” Jurnal Informatika, vol. 4, no. 2, pp. 112–119, 2019.

[2] Berry, M. J., & Linoff, G., Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. Wiley, 2004.

[3] Thomas, L. C., Consumer Credit Models: Pricing, Profit and Portfolios. Oxford University Press, 2009.

[4] Han, J., & Kamber, M., Data Mining: Concepts and Techniques. Elsevier, 2006.

[5] Einav, L., Jenkins, M., & Levin, J., “Contract pricing in consumer credit markets,” Econometrica, vol. 81, no. 4, pp. 123–134, 2013.

[6] Acock, A. C., A Gentle Introduction to Stata. Stata Press, 2005.

[7] Osama, M., “Performance analysis of EM and K-Means clustering algorithms,” International Journal of Computer Applications, vol. 45, no. 3, pp. 21–26, 2008.

[8] Tengke Xiong, et al., “Data clustering for credit risk assessment,” Journal of Financial Risk Management, vol. 5, no. 1, pp. 34–47, 2013.

[9] Lee, T., & Chen, H., “Hybrid models for credit scoring,” Decision Support Systems, vol. 42, no. 1, pp. 16–24, 2005.

[10] Huang, C., Chen, M., & Wang, C., “A hybrid method for credit risk prediction,” Expert Systems with Applications, vol. 27, no. 4, pp. 60–70, 2004.

[11] Xiong, T., “Clustering methods in financial data mining,” Computational Economics, vol. 20, no. 3, pp. 225–238, 2013.

[12] Ian, H., & Iebe, F., “The role of EM in modern data analysis,” Journal of Data Science, vol. 5, no. 1, pp. 67–81, 2005.

[13] W. Piatetsky-Shapiro, "CRISP-DM: Towards a Standard Process Model for Data Mining," in *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, London, 2000.

[14] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Waltham: Elsevier, 2012.

[15] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer, 2009.

[16] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. San Francisco: Morgan Kaufmann, 2016.

[17] R. Kohavi and F. Provost, "Glossary of Terms for Evaluating Machine Learning Algorithms," *Machine Learning*, vol. 30, no. 2-3, pp. 271–274, 1998.

[18] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Magazine*, vol. 17, no. 3, pp. 37–54, 1996.