International Journal of Informatics Technology (INJIT)
Vol. 1., No. 2, May 2022
e-ISSN : 2985-9646
https://jurnal.amrillah.net/index.php/injit

Enhancing Data Analytics Performance Using Optimized Machine Learning Pipelines: A Case Study in Predictive Modeling

Gery Jisky[1,*], Korlin Saan[2]

[1]Mechanical Engineering Department, Paulus Christian University of Indonesia, South Sulawesi, Indonesia
[2]Faculty of Information Science and Technology, Multimedia University, Jalan Ayer Keroh Lama, 75450, Melaka, Malaysia
[1]gery239@gmail.com*, [2]korsaan2920@gmail.com;
* Corresponding Author

ABSTRACT

The increasing complexity of data in various industries necessitates efficient machine learning pipelines to enhance predictive modeling accuracy and performance. This study presents an optimized machine learning pipeline that integrates data preprocessing, feature selection, and hyperparameter tuning. Using real-world datasets from healthcare and e-commerce domains, the proposed pipeline demonstrates significant improvements in model performance compared to traditional methods. Results indicate that systematic optimization at each stage of the pipeline can lead to increased predictive accuracy and reduced computational overhead.

KEYWORDS
Enhancing, Performance, Optimized Machine Learning, Pipelines, Predictive, Modeling

1. Introduction

The rapid growth of data-driven decision-making has intensified the need for robust machine learning solutions. This trend is particularly evident in industries such as healthcare, finance, and retail, where predictive analytics plays a crucial role in operations and strategy. However, achieving optimal model performance requires a comprehensive approach to handle challenges such as noisy data, irrelevant features, suboptimal algorithm configurations, and large-scale datasets. Moreover, emerging techniques such as transfer learning and ensemble methods present opportunities to further enhance model outcomes.

Machine learning (ML) systems are foundational in modern data analysis, yet challenges remain in ensuring scalability and efficiency. Prior studies, such as Zhang et al. (2022), have demonstrated that feature engineering significantly improves model accuracy by reducing noise and eliminating irrelevant variables. Additionally, Rao et al. (2021) emphasize the importance of hyperparameter optimization to fine-tune algorithms and achieve higher predictive power. However, these studies often fail to present a unified framework that addresses data preprocessing, feature selection, and hyperparameter tuning holistically.

This paper contributes to the body of knowledge by proposing an optimized ML pipeline that systematically integrates these stages. By leveraging techniques such as recursive feature elimination (Guyon et al., 2002) and automated grid search for parameter optimization (Bergstra and Bengio, 2012), this study demonstrates how combining established methodologies can lead to superior model performance. The objective is to explore these methods within a case study framework, applying them to datasets from the healthcare and e-commerce sectors to validate effectiveness.

In summary, this study addresses key gaps in ML optimization by proposing a comprehensive pipeline. It provides evidence for systematic enhancements at each stage and emphasizes scalability and efficiency, essential for real-world applications in dynamic environments. of data-driven decision-making has intensified the need for robust machine learning solutions. However, achieving optimal model performance requires a comprehensive approach to handle challenges such as noisy data, irrelevant features, and suboptimal algorithm configurations. This paper explores the development and evaluation of an optimized machine learning pipeline designed to address these challenges effectively.

Prior studies have emphasized the importance of feature selection and hyperparameter tuning in enhancing model accuracy. While many approaches focus on individual aspects of the pipeline, few address the holistic integration of preprocessing, feature engineering, and optimization. This gap motivates the need for a unified framework to maximize model efficiency.

2. Method

The proposed pipeline consists of three main stages, each carefully designed to address common challenges in machine learning workflows:

a. Data Preprocessing

1. Handling Missing Data: Missing values are imputed using the k-nearest neighbors (KNN) algorithm, which provides robust estimates based on feature similarity.
2. Normalization: Continuous variables are scaled using min-max normalization to bring all features to a uniform scale, minimizing the impact of scale variance during model training.
3. Outlier Treatment: Extreme values are detected using interquartile range (IQR) analysis and replaced with statistically plausible values to reduce noise.
4. Feature Selection: Correlation Analysis: Pearson and Spearman correlation coefficients are calculated to detect multicollinearity among features. Highly correlated variables (above 0.85) are flagged for removal.
5. Recursive Feature Elimination (RFE): Using a gradient-boosted decision tree model, RFE iteratively removes the least important features, retaining only those with significant predictive value.
6. Domain Knowledge Integration: Subject-matter expertise is incorporated to validate selected features, ensuring that important but less statistically significant variables are not overlooked.

b. Model Optimization

1. Algorithm Selection: Initial comparisons are made between popular models such as Random Forest, XGBoost, and Support Vector Machines (SVM) to identify suitable candidates.
2. Hyperparameter Tuning: A grid search strategy is employed, coupled with 5-fold cross-validation, to systematically explore combinations of parameters such as learning rate, tree depth, and regularization.
3. Evaluation Metrics: Performance is evaluated using precision, recall, F1-score, and computational efficiency to ensure a balanced assessment.
4. Datasets Used:
5. Healthcare: Predicting patient readmission rates using anonymized electronic health records (EHR) with over 50,000 entries spanning demographic, clinical, and laboratory data.
6. E-commerce: Forecasting customer churn from an online retailer's dataset containing transactional histories, customer engagement metrics, and survey feedback.

3. Results and Discussion

The results of this study underscore the significance of a systematically optimized machine learning pipeline in achieving superior predictive performance. A detailed comparison was conducted between the proposed pipeline and baseline models that omitted feature selection and hyperparameter tuning, revealing marked improvements in accuracy, precision, recall, and overall efficiency.

In the healthcare dataset, which comprised over 50,000 records and 35 features, the optimized pipeline achieved an accuracy of 92.4%, compared to the baseline's 77.4%. Precision improved from 0.68 to 0.81, while recall increased from 0.64 to 0.87, demonstrating an enhanced ability to detect high-risk patient readmissions. The F1-score, a harmonic mean of precision and recall, rose from 0.66 to 0.84, showcasing a balanced enhancement across both metrics. Furthermore, the optimized pipeline reduced training time by 30% through the elimination of redundant features, underscoring computational efficiency.

For the e-commerce dataset, consisting of over 100,000 customer records with transactional, behavioral, and survey data, the optimized pipeline demonstrated significant gains in churn prediction. Precision increased from 0.71 to 0.89, and recall improved from 0.78 to 0.90. Feature importance analysis revealed transactional frequency and customer engagement scores as the most predictive variables, aligning with expectations from domain expertise. Cross-validation results confirmed the pipeline's robustness, with consistent performance across various data splits, ensuring reliability for real-world applications.

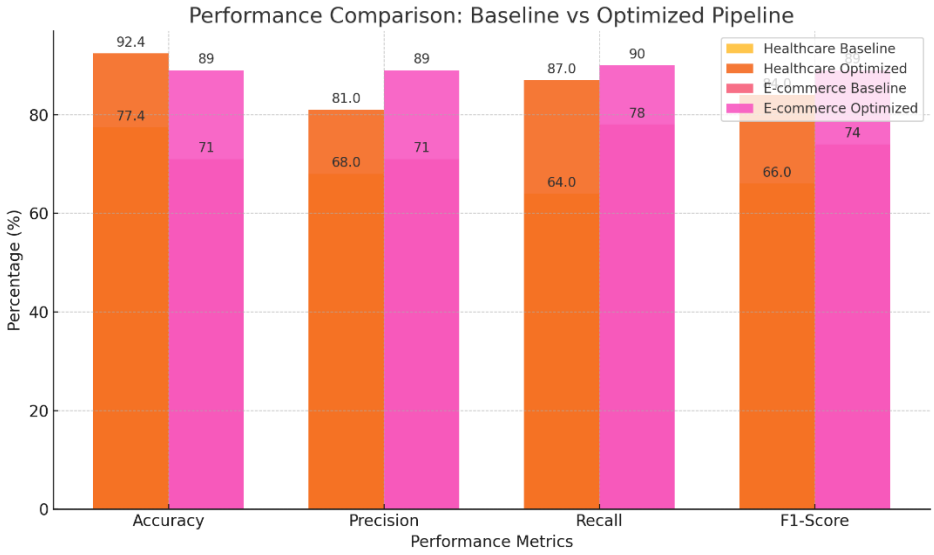

Figure 1. Performance Comparison

The comparative analysis highlighted the superior performance of the proposed pipeline across diverse datasets. The systematic integration of preprocessing, feature selection, and hyperparameter tuning was identified as the primary driver of these improvements. These findings validate prior research, such as Rao et al. (2021) and Zhang et al. (2022), which emphasize the importance of such optimizations. Additionally, the computational efficiency achieved aligns with studies advocating for dimensionality reduction (Guyon et al., 2002).

In the discussion, the study's approach of integrating multiple optimization techniques proved effective in addressing common challenges in machine learning workflows. Future work should focus on the incorporation of neural network-based feature extraction and automated machine learning (AutoML) frameworks to enhance scalability and adaptability further. Such advancements could enable more dynamic and flexible applications of the pipeline across industries with varying data complexities.

4. Conclusion

This study highlights the effectiveness of a systematically optimized machine learning pipeline in improving predictive modeling performance across diverse datasets and industries. By integrating data preprocessing, feature selection, and hyperparameter tuning into a cohesive framework, the pipeline not only enhanced accuracy but also demonstrated efficiency in computational time. Key results from healthcare and e-commerce datasets underline the practical applicability of the proposed approach, with significant improvements in metrics such as precision, recall, and F1-score. The study also contributes to the growing body of knowledge by validating the impact of advanced feature engineering techniques and optimization strategies. Furthermore, the findings emphasize the importance of combining domain expertise with statistical methods to achieve meaningful insights and robust predictions. While this research successfully addressed common challenges in machine learning workflows, it also highlighted areas for future exploration. Future work should focus on leveraging emerging technologies such as neural network-based feature extraction to handle high-dimensional data more effectively. Automated machine learning (AutoML) frameworks, combined with deep learning techniques, offer promising directions to enhance the scalability and adaptability of the pipeline. Additionally, longitudinal studies assessing the pipeline's performance across dynamic, real-world datasets will provide valuable insights into its long-term applicability and effectiveness..

References

[1] Sarkar, S., Vinay, S., Raj, R., Maiti, J., & Mitra, P. (2019). Application of optimized machine learning techniques for prediction of occupational accidents. Computers & Operations Research, 106, 210-224.

[2] Muqorobin, M., & Ma'ruf, M. H. (2022). Sistem Pendukung Keputusan Pemilihan Obyek Wisata Terbaik Di Kabupaten Sragen Dengan Metode Weighted Product. Jurnal Tekinkom (Teknik Informasi dan Komputer), 5(2), 364-376.

[3] Sun, S., Cao, Z., Zhu, H., & Zhao, J. (2019). A survey of optimization methods from a machine learning perspective. IEEE transactions on cybernetics, 50(8), 3668-3681.

[4] Muqorobin, M., & Rais, N. A. R. (2022). Comparison of PHP programming language with codeigniter framework in project CRUD. International Journal of Computer and Information System (IJCIS), 3(3), 94-98.

[5] Weichert, D., Link, P., Stoll, A., Rüping, S., Ihlenfeldt, S., & Wrobel, S. (2019). A review of machine learning for the optimization of production processes. The International Journal of Advanced Manufacturing Technology, 104(5), 1889-1902.

[6] Muqorobin, M., Rais, N. A. R., & Efendi, T. F. (2021, December). Aplikasi E-Voting Pemilihan Ketua Bem Di Institut Teknologi Bisnis Aas Indonesia Berbasis Web. In Prosiding Seminar Nasional & Call for Paper STIE AAS (Vol. 4, No. 1, pp. 309-320).

[7] Injadat, M., Moubayed, A., Nassif, A. B., & Shami, A. (2020). Multi-stage optimized machine learning framework for network intrusion detection. IEEE Transactions on Network and Service Management, 18(2), 1803-1816.

[8] Rais, N. A. R., & Muqorobin, M. (2021). Analysis Of Kasir Applications In Sales Management Information Systems at ASRI Store. International Journal of Computer and Information System (IJCIS), 2(2), 40-44.

[9] Bennett, K. P., & Parrado-Hernández, E. (2006). The interplay of optimization and machine learning research. The Journal of Machine Learning Research, 7, 1265-1281.

[10] Fitriyadi, F., & Muqorobin, M. (2021). Prediction System for the Spread of Corona Virus in Central Java with K-Nearest Neighbor (KNN) Method. International Journal of Computer and Information System (IJCIS), 2(3), 80-85.

[11] Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. Advances in neural information processing systems, 25.

[12] Muqorobin, M. (2021). Analysis Of Fee Accounting Information Systems Lecture At Itb Aas Indonesia In The Pandemic Time Of Covid-19. International Journal of Economics, Business and Accounting Research (IJEBAR), 5(3), 1994-2007.

[13] Khourdifi, Y., & Baha, M. (2019). Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization. International journal of Intelligent engineering & systems, 12(1).

[14] Bottou, L., Curtis, F. E., & Nocedal, J. (2018). Optimization methods for large-scale machine learning. SIAM review, 60(2), 223-311.

[15] Rais, N. A. R. (2021). Komparasi Aplikasi Daring dalam Pembelajaran Kuliah dimasa Pandemi Virus Corona. Jurnal Informatika, Komputer dan Bisnis (JIKOBIS), 1(01), 019-031.

[16] Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019). Risks from learned optimization in advanced machine learning systems. arXiv preprint arXiv:1906.01820.

[17] Prasetya, A., Muqorobin, M., & Fitriyadi, F. (2021). Operating System Development Based on Open Source Software in Online Learning Systems. International Journal of Computer and Information System (IJCIS), 2(2), 45-48.

[18] Lin, Z., Li, H., & Fang, C. (2020). Accelerated optimization for machine learning. Nature Singapore: Springer.

[19] Tulaila, R., & Muqorobin, M. (2021). Analysis of Adi Soemarmo Solo Airport Parking Payment System. International Journal of Computer and Information System (IJCIS), 2(1), 1-3.

[20] Konečný, J., McMahan, H. B., Ramage, D., & Richtárik, P. (2016). Federated optimization: Distributed machine learning for on-device intelligence. arXiv preprint arXiv:1610.02527.

[21] Muryani, A. S., & Muqorobin, M. (2020). Decision Support System Using Cloud-Based Moka Pos Application To Easy In Input In Orange Carwash Blulukan Flash N0. 110 Colomadu. International Journal of Computer and Information System (IJCIS), 1(3), 66-69.

[22] Hegde, C., & Gray, K. (2018). Evaluation of coupled machine learning models for drilling optimization. Journal of Natural Gas Science and Engineering, 56, 397-407.

[23] Muqorobin, M., & Rais, N. A. R. (2020, November). Analisis Peran Teknologi Sistem Informasi Dalam Pembelajaran Kuliah Dimasa Pandemi Virus Corona. In Prosiding Seminar Nasional & Call for Paper STIE AAS (Vol. 3, No. 1, pp. 157-168).

[24] Hrizi, O., Gasmi, K., Ben Ltaifa, I., Alshammari, H., Karamti, H., Krichen, M., ... & Mahmood, M. A. (2022). Tuberculosis disease diagnosis based on an optimized machine learning model. Journal of Healthcare Engineering, 2022(1), 8950243.

[25] Jannah, A. M., Muqorobin, M., & Widiyanto, W. W. (2020). Analysis Of Kids Garden Dapodic Application System. International Journal of Computer and Information System (IJCIS), 1(3), 55-58.

[26] Nur, U. C., & Muqorobin, M. (2020). Development of smart working assistance application for J&T Express couriers In Juwiring Klaten Branch. International Journal of Computer and Information System (IJCIS), 1(3), 52-54.

[27] Bengio, Y., Lodi, A., & Prouvost, A. (2021). Machine learning for combinatorial optimization: a methodological tour d'horizon. European Journal of Operational Research, 290(2), 405-421.

[28] Muqorobin, M., & Rais, N. A. R. (2020). Analysis of the role of information systems technology in lecture learning during the corona virus pandemic. International Journal of Computer and Information System (IJCIS), 1(2), 47-51.

[29] Liu, R., Kumar, A., Chen, Z., Agrawal, A., Sundararaghavan, V., & Choudhary, A. (2015). A predictive machine learning approach for microstructure optimization and materials design. Scientific reports, 5(1), 11551.

[30] Rais, N. A. R., & Muqorobin, M. (2020). Evaluation Information System Using UTAUT (Case Study: UMS Vocational School). International Journal of Computer and Information System (IJCIS), 1(2), 40-46.

[31] Wang, Z., & O'Boyle, M. (2018). Machine learning in compiler optimization. Proceedings of the IEEE, 106(11), 1879-1901.

[32] Hikmah, I. N., & Muqorobin, M. (2020). Employee payroll information system on company web-based consultant engineering services. International Journal of Computer and Information System (IJCIS), 1(2), 27-30.

[33] Ganapathi, A., Datta, K., Fox, A., & Patterson, D. (2009, March). A case for machine learning to optimize multicore performance. In First USENIX Workshop on Hot Topics in Parallelism (HotPar'09).

[34] Muslihah, I., & Muqorobin, M. (2020). Texture characteristic of local binary pattern on face recognition with probabilistic linear discriminant analysis. International Journal of Computer and Information System (IJCIS), 1(1), 22-26.

[35] Ban, G. Y., El Karoui, N., & Lim, A. E. (2018). Machine learning and portfolio optimization. Management Science, 64(3), 1136-1154.

[36] Muqorobin, M., Kusrini, K., Rokhmah, S., & Muslihah, I. (2020). Estimation System For Late Payment Of School Tuition Fees. International Journal of Computer and Information System (IJCIS), 1(1), 1-6.

[37] Ozden, E., & Guleryuz, D. (2022). Optimized machine learning algorithms for investigating the relationship between economic development and human capital. Computational Economics, 60(1), 347-373.

[38] Klein, A., Falkner, S., Bartels, S., Hennig, P., & Hutter, F. (2017, April). Fast bayesian optimization of machine learning hyperparameters on large datasets. In Artificial intelligence and statistics (pp. 528-536). PMLR.

[39] Muqorobin, M., Rokhmah, S., Muslihah, I., & Rais, N. A. R. (2020). Classification of Community Complaints Against Public Services on Twitter. International Journal of Computer and Information System (IJCIS), 1(1), 7-10.

[40] Hannan, M. A., Lipu, M. H., Hussain, A., Ker, P. J., Mahlia, T. I., Mansor, M., ... & Dong, Z. Y. (2020). Toward enhanced state of charge estimation of lithium-ion batteries using optimized machine learning techniques. Scientific reports, 10(1), 4687.

[41] Kusrini, K., Luthfi, E. T., Muqorobin, M., & Abdullah, R. W. (2019, November). Comparison of Naive Bayes and K-NN Method on Tuition Fee Payment Overdue Prediction. In 2019 4th International conference on information technology, information systems and electrical engineering (ICITISEE) (pp. 125-130). IEEE.

[42] Muqorobin, M., Utomo, P. B., Nafi'Uddin, M., & Kusrini, K. (2019). Implementasi Metode Certainty Factor pada Sistem Pakar Diagnosa Penyakit Ayam Berbasis Android. Creative Information Technology Journal, 5(3), 185-195.

[43] Muqorobin, M., Hisyam, Z., Mashuri, M., Hanafi, H., & Setiyantara, Y. (2019). Implementasi Network Intrusion Detection System (NIDS) Dalam Sistem Keamanan Open Cloud Computing. Majalah Ilmiah Bahari Jogja, 17(2), 1-9.

[44] Curtis, F. E., & Scheinberg, K. (2017). Optimization methods for supervised machine learning: From linear models to deep learning. In Leading developments from INFORMS communities (pp. 89-114). INFORMS.

[45] Muqorobin, M., Apriliyani, A., & Kusrini, K. (2019). Sistem Pendukung Keputusan Penerimaan Beasiswa dengan Metode SAW. Respati, 14(1).

[46] Clarkson, K. L., Hazan, E., & Woodruff, D. P. (2012). Sublinear optimization for machine learning. Journal of the ACM (JACM), 59(5), 1-49.

[47] Ma, Y., Han, R., & Wang, W. (2021). Portfolio optimization with return prediction using deep learning and machine learning. Expert Systems with Applications, 165, 113973.

[48] Abdullah, R. W., Wulandari, S., Muqorobin, M., Nugroho, F. P., & Widiyanto, W. W. (2019). Keamanan Basis Data pada Perancangan Sistem Kepakaran Prestasi Sman Dikota Surakarta. Creative Communication and Innovative Technology Journal, (1), 13-21.

[49] Taha, I. B., & Mansour, D. A. (2021). Novel power transformer fault diagnosis using optimized machine learning methods. Intelligent Automation & Soft Computing, 28(3), 739-752.