Jurnal Informatika Ekonomi Bisnis

http://www.infeb.org

2024 Vol. 6 Iss. 1 Hal: 246-250 e-ISSN: 2714-8491

Optimizing Machine Learning Models for Urinary Tract Infection Diagnostics: A Comparative Study of Logistic Regression and Random Forest

Gregorius Airlangga[1]✉

[1]Atma Jaya Catholic University of Indonesia

gregorius.airlangga@atmajaya.ac.id

Abstract

Urinary Tract Infections (UTIs) present a significant healthcare challenge due to their prevalence and diagnostic complexity. Timely and accurate diagnosis is critical for effective treatment, yet traditional methods like microbial cultures and urinalysis are often slow and inconsistent. This study introduces machine learning (ML) as a transformative solution for UTI diagnostics, particularly focusing on logistic regression and random forest models renowned for their interpretability and robustness. We conducted a meticulous hyperparameter tuning process using a rich dataset from a clinic in Northern Mindanao, Philippines, incorporating demographic, clinical, and urinalysis data. Our research outlines a detailed methodology for applying and refining these ML models to predict UTI outcomes accurately. Through comprehensive hyperparameter optimization, we enhanced the predictive performance, demonstrating a significant improvement over standard diagnostic practice. The findings reveal a clear superiority of the random forest model, achieving a top testing accuracy of 0.9814, compared to the best-performing logistic regression model's accuracy of 0.7626. This comparative analysis not only validates the efficacy of ML in medical diagnostics but also emphasizes the potential clinical impact of these models in real-world settings. The study contributes to the burgeoning literature on ML applications in healthcare by providing a blueprint for optimizing ML models for clinical use, particularly in diagnosing UTIs. It underscores the promise of ML in augmenting diagnostic precision, thereby potentially reducing the global healthcare burden associated with UTIs.

Keywords: Logistic Regression, Random Forest, UTI Detection, UTI Diagnostics, Machine Learning.

1. Introduction

Urinary Tract Infections (UTIs) are among the most common bacterial infections, afflicting individuals across all age brackets and contributing significantly to the global healthcare burden [1], [2], [3]. The complexity of UTI diagnosis, influenced by a broad spectrum of symptoms and causative pathogens, underscores the need for more precise, efficient, and rapid diagnostic methods [4], [5], [6]. While traditional diagnostic techniques rely on microbial cultures and urinalysis, these methods often suffer from delays and variable sensitivity, which can hinder timely and accurate treatment [7], [8], [9].

The burgeoning field of machine learning (ML) offers promising tools for revolutionizing UTI diagnostics [10], [11], [12]. Among the plethora of ML methodologies, logistic regression and random forest models stand out for their interpretability, efficiency, and robust performance across various predictive modeling tasks [13], [14], [15]. These models can integrate and analyze complex, multidimensional datasets, offering insights and predictive capabilities far beyond traditional statistical approaches [16], [17], [18]. Despite their potential, the optimal application of these models in UTI diagnostics requires careful tuning and validation to ensure maximum accuracy and utility in real-world settings [19], [20], [21].

Recent literature has begun to explore the application of ML models in diagnosing UTIs, demonstrating their potential to enhance diagnostic precision and efficiency [22], [23], [24]. However, a significant gap remains in the optimization of these models for clinical use. Many studies fail to thoroughly explore hyperparameter tuning, a critical process for enhancing model performance [25], [26], [27]. Moreover, there's a need for research that not only demonstrates the efficacy of ML models but also details the methodology for achieving optimal model configuration, specifically in the context of UTI diagnostics [28], [29], [30].

This study seeks to fill these gaps by focusing on the application and meticulous hyperparameter tuning of logistic regression and random forest models for UTI diagnosis. Utilizing a comprehensive dataset collected from a local clinic in Northern Mindanao, Philippines, this research encompasses a wide array of variables pertinent to UTI diagnosis, including demographic, clinical, and urinalysis data [31]. The dataset's richness allows for a nuanced exploration of model performance in predicting UTI outcomes, offering a robust testing ground for our ML models.

Accepted: 22-03-2024 | Revision: 27-03-2024 | Publication: 31-03-2024 | doi: 10.37034/infeb.v6i1.854

The primary contributions of this research are twofold. First, we present a detailed methodology for the application and optimization of logistic regression and random forest models in UTI diagnostics, highlighting the importance of hyperparameter tuning in maximizing model performance. Second, our study provides a comparative analysis of these models, offering insights into their relative strengths and limitations in the context of UTI diagnosis. Through rigorous validation and optimization, this research aims to advance the practical application of ML in enhancing diagnostic accuracy for UTIs.

The article is structured to guide the reader through the entire research process. Following this introduction, Section 2 elaborates on the materials and methods, including dataset description, model selection rationale, and the hyperparameter tuning approach. Section 3 presents the results, showcasing the performance of logistic regression and random forest models post-optimization. Finally, Section 4 concludes the article, summarizing the key insights and contributions.

2. Research Method

2.1 Dataset Description

This study utilizes a comprehensive dataset collected from a local clinic in Northern Mindanao, Philippines, spanning from April 2020 to January 2023 [31]. The dataset comprises patient records associated with urinalysis tests, crucial for diagnosing Urinary Tract Infections (UTIs). It includes demographic information (age and gender), urine physical characteristics (color and transparency), chemical properties (glucose, protein, pH, specific gravity), and microscopic examination findings (white blood cells (WBC), red blood cells (RBC), epithelial cells, mucous threads, amorphous urates, and bacteria presence). The target variable, UTI diagnosis, is binary, indicating the presence or absence of infection. The dataset's diversity and comprehensiveness enable the exploration of machine learning models to predict UTI outcomes effectively.

2.2 Preprocessing and Feature Engineering

The initial step in the analysis involved filtering warnings to ensure clarity in presenting the results. Data preprocessing included handling both nominal and ordinal features distinctively. The ordinal features, such as Transparency, Epithelial Cells, Mucous Threads, Amorphous Urates, Bacteria, Color, Protein, Glucose, WBC, and RBC, were encoded based on their inherent order. A custom sorting algorithm was applied to WBC and RBC features to handle various data representations, including ranges, greater-than signs, and textual descriptors like "LOADED" and "TNTC" (Too Numerous to Count), with a systematic approach to maintain their ordinality. Then, Nominal features like Gender were encoded using One-Hot Encoding to convert categories into a binary vector format. For the critical features of WBC and RBC, a novel binning process was employed to manage the high cardinality stemming from unique values. This process involved creating bins based on sorted values, reducing dimensionality while preserving the ordinal nature and valuable information within these features.

Furthermore, Data normalization, particularly for continuous features, was performed using MinMaxScaler, adjusting feature scales to a common range, enhancing model training efficiency. SMOTE (Synthetic Minority Over-sampling Technique) was utilized to address class imbalance, generating synthetic samples to ensure balanced representation of both classes in the target variable.

2.3 Model Selection and Rationale

The research focuses on Logistic Regression and Random Forest models, selected for their robustness, interpretability, and widespread application in binary classification problems. Logistic Regression, a linear model, is chosen for its simplicity and efficacy in estimating probabilities, providing a solid baseline for performance comparison. The Random Forest model, known for handling non-linear relationships and feature interactions, offers a contrast to Logistic Regression with its ensemble learning approach, potentially capturing complex patterns in the data.

2.4 Hyperparameter Tuning Approach

Optuna, an open-source hyperparameter optimization framework, was employed for tuning the models. This framework is preferred for its efficiency in searching through the hyperparameter space using a Bayesian optimization approach, significantly reducing the computational expense compared to traditional grid search methods. Each model underwent a series of 100 trials, with Optuna tasked to maximize the F1-score, a metric chosen for its balance between precision and recall, critical in the context of imbalanced datasets like UTI diagnostics.

The hyperparameter tuning for Logistic Regression involved optimizing the tolerance for stopping criteria ('tol') and the inverse of regularization strength ('C'), along with a threshold for classifying positive instances. For the Random Forest model, the number of trees ('n_estimators'), the maximum depth of the trees ('max_depth'), the number of features to consider for the best split ('max_features'), and the criterion for measuring the quality of a split ('criterion') were tuned.

2.5 Implementation Details

The analysis was implemented using Python, leveraging libraries such as Pandas for data manipulation, Sklearn for modeling and preprocessing tools, Imblearn for oversampling, and Matplotlib and Seaborn for visualization. The code was executed in a Google Colab environment, facilitating access to high computational resources and a collaborative platform for the research team.

2.6 Evaluation Metrics

Model performance was primarily evaluated using the F1-score, considering the dataset's imbalance. Additionally, accuracy, recall, precision, and the Area Under the Receiver Operating Characteristic (ROC AUC) score were used to provide a comprehensive assessment of each model's predictive capabilities.

3. Result and Discussion

The top 10 logistic regression models can be seen on Table 1.

Table 1. Testing Accuracy and Configuration of Hyper Parameters for Logistic Regression

Testing Accuracy	C	Threshold	Tolerance
0.7626	0.6019401452459415	0.28564893809036207	1.1407041351358772e-05
0.7626	0.5797450437266929	0.28884705567532100	2.5915178745198782e-06
0.7626	0.5808107586143879	0.28698585090769957	5.2982956659526150e-06
0.7626	0.5630515096981310	0.28493445329888140	2.5664995486634457e-06
0.7626	0.5128480830956532	0.28791700530780860	2.6381293477262452e-06
0.7626	0.5606516733016459	0.29011186581790477	4.0008549271281250e-06
0.7626	0.5582276602243045	0.28619462227526240	2.6593131170835832e-06
0.7626	0.5592175461842079	0.28300269848793497	2.1474620014371140e-06
0.7626	0.5314887994518090	0.28540356367893030	0.00014829870333910985
0.7626	0.6190594942360033	0.29182946968101540	0.00013316398716669225

The top 10 logistic regression models, as presented in Table 1, all achieved a testing accuracy of 0.7626. This uniformity in performance indicates that despite the variations in hyperparameters, the models were able to achieve the same level of accuracy on the test set. The regularization strength (C) among the top logistic regression models varied, with values ranging from approximately 0.512 to 0.621, which suggests a level of robustness against overfitting without significant impact on the accuracy. The threshold, which is likely related to the decision function for class separation, shows minor variations across models, all hovering around 0.28. This points to a consistent classification boundary being determined by the logistic regression models. Furthermore, the tolerance for stopping criteria (tol) had wider variations, from 1.14e-05 to almost zero, which did not seem to affect the performance, possibly because the models converged well before the tolerance threshold played a role.

On the other hand, the random forest models can be seen on Table 2.

Table 2. Testing Accuracy and Configuration of Hyper Parameters for Random Forest

TA	Criterion	Max Depth	MF	N Estimator
0.9814	gini	22	2	111
0.9786	gini	22	2	107
0.9760	gini	24	2	119
0.9759	log_loss	30	2	139
0.9734	gini	30	2	100
0.9731	entropy	23	2	135
0.9707	gini	26	Log2	128
0.9705	log_loss	21	2	112
0.9679	log_loss	18	2	92
0.9676	entropy	25	2	107

Where TA is testing accuracy and MF is max features. The random forest models showed a range of testing accuracies from 0.9676 to the best model's accuracy of 0.9814. The variation in accuracy, albeit small, indicates that the random forest's ensemble approach could capture more nuances in the data leading to slightly improved predictions. Then, the best-performing random forest model used a maximum depth of 22, which is indicative of the complexity it could handle, and 111 trees in the ensemble, suggesting a sufficiently diverse set of learners without becoming excessively complex. It's noteworthy that all the top models used 2 features when considering the best split, which demonstrates that only a few features were strong predictors and necessary for making accurate predictions. Furthermore, the n_estimators, which denotes the number of trees in the forest, varied among the top models, yet the accuracies did not change drastically, which could mean that beyond a certain number of trees, the incremental benefit to accuracy diminishes.

Comparing the two types of models, random forest outperformed logistic regression, with the best random forest model achieving an accuracy approximately 2% higher than the logistic regression models. This superiority could be due to the random forest's ability to model non-linear relationships and interactions between features, which logistic regression might not capture as effectively. The hyperparameters for random forest models also displayed a tighter range of values leading to the top accuracies compared to logistic regression, suggesting that fine-tuning is more sensitive for random forest performance. In conclusion, while logistic regression provided a stable solution, random forest demonstrated a higher peak performance, possibly benefiting from its ability to capture more complex patterns in the data. For future modeling endeavors, one might consider exploring more granular adjustments in the random forest's hyperparameters, while for logistic regression, a focus on feature selection and engineering might yield better discriminative power.

4. Conclusion

This investigation into the performance of logistic regression and random forest models within the sphere of UTI diagnostics has illuminated the nuanced capabilities of these machine learning methodologies. Our findings depict logistic regression as a reliable, interpretable, and straightforward model, which achieves a commendable baseline accuracy. Nonetheless, its performance plateau suggests that, within the confines of our dataset and the complex nature of UTI symptoms and pathogens, its capabilities are near their optimization peak. In contrast, the random forest model, with its ensemble-based approach, has demonstrated superior proficiency in managing the intricacies of the data, achieving notable testing accuracy. The significant variance in performance due to hyperparameter adjustments signifies the intricate dance between model complexity and diagnostic precision—a balance that is crucial in the medical field for actionable insights. The study reinforces that the logistic regression model, despite its transparency, may be less suited for the multifaceted patterns present in UTI data, while the random forest model is more adept at navigating through these complexities, albeit at the cost of interpretability and increased computational demand. The choice between these models should be informed by the specific requirements of the diagnostic challenge, weighing the trade-offs between simplicity and performance, interpretability and computational intensity.

References

[1] Öztürk, R., & Murt, A. (2020). Epidemiology of urological infections: a global burden. *World journal of urology*, *38*, 2669-2679. https://doi.org/10.1007/s00345-019-03071-4

[2] Akhtar, A., Ahmad Hassali, M. A., Zainal, H., & Khan, A. H. (2021). A cross-sectional assessment of urinary tract infections among geriatric patients: prevalence, medication regimen complexity, and factors associated with treatment outcomes. *Frontiers in public health*, *9*, 657199.

[3] Mancuso, G., Midiri, A., Gerace, E., Marra, M., Zummo, S., & Biondo, C. (2023). Urinary tract infections: the current scenario and future prospects. *Pathogens*, *12*(4), 623. https://doi.org/10.3390/pathogens12040623

[4] Iriya, R., Braswell, B., Mo, M., Zhang, F., Haydel, S. E., & Wang, S. (2024). Deep Learning-Based Culture-Free Bacteria Detection in Urine Using Large-Volume Microscopy. *Biosensors*, *14*(2), 89. https://doi.org/10.3390/bios14020089

[5] Schinas, G., Dimopoulos, G., & Akinosoglou, K. (2023). Understanding and implementing diagnostic stewardship: a guide for resident physicians in the era of antimicrobial resistance. *Microorganisms*, *11*(9), 2214. https://doi.org/10.3390/microorganisms11092214

[6] Hasan, J., & Bok, S. (2024). Plasmonic Fluorescence Sensors in Diagnosis of Infectious Diseases. *Biosensors*, *14*(3), 130. https://doi.org/10.3390/bios14030130

[7] Jarzembowski, T., & Daca, A. (Eds.). (2024). Advances and Challenges in Urine Laboratory Analysis.

[8] Kight, E. C. (2023). *Advancements in Point-of-Care Diagnostic Assays for Non-Invasive Samples in Resource Constrained Settings* (Doctoral dissertation, Vanderbilt University).

[9] Sykes, J. E., Reagan, K. L., Nally, J. E., Galloway, R. L., & Haake, D. A. (2022). Role of diagnostics in epidemiology, management, surveillance, and control of leptospirosis. *Pathogens*, *11*(4), 395. https://doi.org/10.3390/pathogens11040395

[10] Kumar, Y., Koul, A., Sisodia, P. S., Shafi, J., Kavita, V., Gheisari, M., & Davoodi, M. B. (2021). Heart failure detection using quantum-enhanced machine learning and traditional machine learning techniques for internet of artificially intelligent medical things. *Wireless Communications and Mobile Computing*, *2021*, 1-16. https://doi.org/10.1155/2021/1616725

[11] Tsai, A. Y., Carter, S. R., & Greene, A. C. (2024, January). Artificial Intelligence in Pediatric Surgery. In *Seminars in Pediatric Surgery* (p. 151390). WB Saunders. https://doi.org/10.1016/j.sempedsurg.2024.151390

[12] Pachiyannan, P., Alsulami, M., Alsadie, D., Saudagar, A. K. J., AlKhathami, M., & Poonia, R. C. (2024). A Novel Machine Learning-Based Prediction Method for Early Detection and Diagnosis of Congenital Heart Disease Using ECG Signal Processing. *Technologies*, *12*(1), 4. https://doi.org/10.3390/technologies12010004

[13] Efron, B. (2020). Prediction, estimation, and attribution. *International Statistical Review*, *88*, S28-S59. https://doi.org/10.1111/insr.12409

[14] Dastile, X., Celik, T., & Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, *91*, 106263. https://doi.org/10.1016/j.asoc.2020.106263

[15] Zhao, B., Song, R., Guo, X., & Yu, L. (2023). Bridging Interpretability and Performance: Enhanced Machine Learning-based Prediction of Hematoma Expansion Post-Stroke via Comprehensive Feature Selection. *IEEE Access*. https://doi.org/10.1109/ACCESS.2023.3348244

[16] Rajula, H. S. R., Verlato, G., Manchia, M., Antonucci, N., & Fanos, V. (2020). Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. *Medicina*, *56*(9), 455. https://doi.org/10.3390/medicina56090455

[17] Shu, X., & Ye, Y. (2023). Knowledge Discovery: Methods from data mining and machine learning. *Social Science Research*, *110*, 102817. https://doi.org/10.1016/j.ssresearch.2022.102817

[18] Mannering, F., Bhat, C. R., Shankar, V., & Abdel-Aty, M. (2020). Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. *Analytic methods in accident research*, *25*, 100113. https://doi.org/10.1016/j.amar.2020.100113

[19] De Bruyne, S., De Kesel, P., & Oyaert, M. (2023). Applications of Artificial Intelligence in Urinalysis: Is the Future Already Here? *Clinical Chemistry*, *69*(12), 1348-1360. https://doi.org/10.1093/clinchem/hvad136

[20] Caruccio, L., Cirillo, S., Polese, G., Solimando, G., Sundaramurthy, S., & Tortora, G. (2024). Can ChatGPT provide intelligent diagnoses? A comparative study between predictive models and ChatGPT to define a new medical diagnostic bot. *Expert Systems with Applications*, *235*, 121186. https://doi.org/10.1016/j.eswa.2023.121186

[21] Çubukçu, H. C., Topcu, D. İ., & Yenice, S. (2023). Machine learning-based clinical decision support using laboratory data. *Clinical Chemistry and Laboratory Medicine (CCLM)*, (0). https://doi.org/10.1515/cclm-2023-1037

[22] Morado, F., & Wong, D. W. (2022). Applying diagnostic stewardship to proactively optimize the management of urinary tract infections. *Antibiotics*, *11*(3), 308. https://doi.org/10.3390/antibiotics11030308

[23] Xu, R., Deebel, N., Casals, R., Dutta, R., & Mirzazadeh, M. (2021). A new gold rush: a review of current and developing diagnostic tools for urinary tract infections. *Diagnostics*, *11*(3), 479. https://doi.org/10.3390/diagnostics11030479

[24] Santos, M., Mariz, M., Tiago, I., Martins, J., Alarico, S., & Ferreira, P. (2022). A review on urinary tract infections diagnostic methods: Laboratory-based and point-of-care approaches. *Journal of Pharmaceutical and Biomedical Analysis*, *219*, 114889. https://doi.org/10.1016/j.jpba.2022.114889

[25] Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., ... & Lindauer, M. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *13*(2), e1484. https://doi.org/10.1002/widm.1484

[26] Asif, D., Bibi, M., Arif, M. S., & Mukheimer, A. (2023). Enhancing heart disease prediction through ensemble learning techniques with hyperparameter optimization. *Algorithms*, *16*(6), 308. https://doi.org/10.3390/a16060308

[27] Quinton, F., Presles, B., Leclerc, S., Nodari, G., Lopez, O., Chevallier, O., ... & Alberini, J. L. (2024). Navigating the nuances: comparative analysis and hyperparameter optimisation of neural architectures on contrast-enhanced MRI for liver and liver tumour segmentation. *Scientific Reports*, *14*(1), 3522. https://doi.org/10.1038/s41598-024-53528-9

[28] Naik, N., Talyshinskii, A., Shetty, D. K., Hameed, B. Z., Zhankina, R., & Somani, B. K. (2024). Smart Diagnosis of Urinary Tract Infections: is Artificial Intelligence the Fast-Lane Solution?. *Current Urology Reports*, *25*(1), 37-47. https://doi.org/10.1007/s11934-023-01192-3

[29] Anahtar, M. N., Yang, J. H., & Kanjilal, S. (2021). Applications of machine learning to the problem of antimicrobial resistance: an emerging model for translational research. *Journal of clinical microbiology*, *59*(7), 10-1128. https://doi.org/10.1128/jcm.01260-20

[30] Luz, C. F., Vollmer, M., Decruyenaere, J., Nijsten, M. W., Glasner, C., & Sinha, B. (2020). Machine learning in infection management using routine electronic health records: tools, techniques, and reporting of future technologies. *Clinical Microbiology and Infection*, *26*(10), 1291-1299. https://doi.org/10.1016/j.cmi.2020.02.003

[31] Avarice02. (2024). Urinalysis Test Results Dataset. Kaggle Dataset. Retrieved from https://www.kaggle.com/datasets/avarice02/urinalysis-test-results