ORIGINAL ARTICLE

Pakistan Journal of Nuclear Medicine
Volume 14(2):49–55
DOI: 10.24911/PJNMed.175-1721068107

Pakistan Journal of Nuclear Medicine is the official journal of Pakistan Society of Nuclear Medicine

Prediction of thyroid cancer recurrence with machine learning models

Dilber Iqbal[1]*, Aamir Shahzad[2]

ABSTRACT

Background: Thyroid cancer recurrence poses a significant challenge in oncology, necessitating effective tools for early prediction. Machine learning (ML) models offer the potential to improve prognostic accuracy and guide clinical decision-making.

Aim and Objective: This study aims to investigate the efficacy of ML models in predicting thyroid cancer recurrence using a publicly available dataset comprising 17 features.

Methods: We explored multiple ML algorithms, including Logistic Regression, K-Nearest Neighbors, Random Forest, and AdaBoost, to develop predictive models. The target variable was the "Recurred" column, indicating whether a patient experienced recurrence. Performance evaluation was conducted using metrics such as Accuracy, Precision, Recall, F1 Score, and receiver operating characteristic (ROC) area under the curve (AUC). A correlation heatmap was generated to assess relationships between features and detect multicollinearity, while feature importance analysis using the Random Forest model identified key predictors.

Results: Among the models, the Random Forest classifier achieved the highest performance on the test dataset, with an Accuracy of 0.9818, Precision of 0.9623, Recall of 1.0000, F1 Score of 0.9808, and ROC AUC of 0.9831. The feature importance analysis highlighted critical factors influencing recurrence prediction, while the correlation heatmap provided insights into feature interactions.

Conclusion: This study demonstrates the effectiveness of ML models, particularly the Random Forest classifier, in predicting thyroid cancer recurrence. The insights gained from feature analysis and correlation studies contribute to model interpretability and future feature selection strategies. These findings emphasize the potential of ML in improving patient outcomes through early and accurate recurrence prediction.

Keywords: Thyroid cancer recurrence, random forest classification model, accuracy, precision, recall, F-1 score.

Received: 30 March 2024 **Accepted:** 26 July 2024

Address for correspondence: Dilber Iqbal
*Department of Medical Physics, PINUM Cancer Hospital, Faisalabad, Pakistan.
Email: dilberiqbal@gmail.com
Full list of author information is available at the end of the article.

Introduction

Thyroid cancer is one of the most common endocrine malignancies, with a generally favorable prognosis. However, despite successful initial treatment, a significant subset of patients experiences recurrence. Recurrence can manifest in various forms, such as local recurrence in the thyroid bed, regional recurrence in the cervical lymph nodes, or distant metastases. Recurrence not only complicates the clinical management of the disease but also impacts the patient's quality of life and survival rates [1,2]. Several factors contribute to the risk of recurrence in thyroid cancer patients, including: Tumor Characteristics: size, histological type, and presence of vascular invasion, aggressive subtypes like poorly differentiated or anaplastic thyroid cancer are more prone to recurrence, Patient Characteristics: age and gender, genetic predispositions and other comorbidities, Initial Treatment: completeness of surgical resection, efficacy of radioactive iodine therapy and thyroid hormone suppression therapy [3-5].

Challenges in Predicting Recurrence

Accurate prediction of thyroid cancer recurrence is challenging due to the heterogeneity of the disease and the complex interplay of various risk factors. Traditional prognostic models, based on clinical and pathological parameters, offer limited predictive power. This necessitates the development of more sophisticated prediction tools to identify high-risk patients who may benefit from closer monitoring and more aggressive adjuvant therapies.

Machine Learning in Recurrence Prediction

Machine learning (ML), a subset of artificial intelligence, provides robust techniques for handling complex, high-dimensional data. ML algorithms can identify subtle patterns and interactions within data that may not be apparent through conventional statistical methods. In the context of thyroid cancer recurrence, ML models can leverage clinical, demographic, and molecular data to develop highly accurate predictive models. There are some Advantages of ML for Recurrence Prediction: High Dimensionality: Ability to process and analyze large datasets with many features, Non-linearity: Capability to model complex, non-linear relationships between features and outcomes and Adaptability: Potential to update and improve models as new data becomes available [3,4,6,7].

ML Classification Models

To predict thyroid cancer recurrence, the diverse array of sophisticated machine-learning classification models exists, each characterized by unique algorithmic paradigms and theoretical foundations [7-9].

Logistic Regression

Logistic Regression operates under the framework of generalized linear models, utilizing a logistic function to model the probability of binary outcomes. It is lauded for its interpretability, as the coefficients provide insight into the relationship between predictor variables and the target outcome, making it a powerful tool for understanding underlying patterns in binary classification tasks.

K-Nearest Neighbors (KNNs)

The Kneighbors Classifier epitomizes a non-parametric, instance-based learning approach. By classifying an observation based on the majority class of its nearest neighbors in the feature space, KNN effectively captures complex, non-linear decision boundaries. Its simplicity belies its efficacy, particularly in contexts where the decision boundary is intricate and difficult to approximate with parametric models.

Random Forest

The Random Forest Classifier embodies the ensemble learning paradigm by constructing a multitude of decision trees during training and aggregating their predictions. This method enhances predictive accuracy and robustness by mitigating overfitting and leveraging the strengths of multiple trees. Each tree is trained on a random subset of the data, a technique known as bootstrap aggregation or bagging, which ensures diversity among the trees and improves generalization [6,10-18].

AdaBoost

The AdaBoost Classifier or Adaptive Boosting is an ensemble technique that sequentially combines weak learners to form a robust classifier. By iteratively adjusting the weights of misclassified instances, AdaBoost focuses subsequent learners on the most challenging cases. This adaptive process enhances the model's capacity to minimize errors and improve overall predictive performance, particularly in scenarios where the underlying data distribution is complex.

These models were meticulously chosen for their complementary strengths and diverse algorithmic foundations, ensuring a comprehensive evaluation of multiple ML methodologies in predicting thyroid cancer recurrence. By leveraging the unique advantages of each model, we aim to identify the most effective approach for this critical clinical prediction task.

Materials and Methods

Dataset

The publicly available CSV dataset used in this study contains a comprehensive set of clinical and demographic features related to thyroid cancer patients [3-5]. The dataset includes various attributes such as the patient's age at diagnosis, gender, current smoking status, smoking history, and history of radiotherapy. It also contains information on thyroid function, results from physical examinations, and the presence of adenopathy. Pathological findings of the tumor, the number of cancer foci, and risk stratification based on clinical parameters are recorded. Additionally, the dataset incorporates tumor size and extent, regional lymph node involvement, and the presence of distant metastasis according to the TNM staging system. The overall cancer stage is based on TNM classification, while the response to initial treatment and whether cancer recurred are also documented, with the latter serving as the target variable indicating recurrence as a binary outcome. This dataset allows for a detailed analysis of various factors (as displayed in Figure 1) influencing thyroid cancer recurrence, enabling the development of robust predictive models.

Limitations of the Study

While the dataset used in this study provides a comprehensive range of clinical and pathological attributes relevant to thyroid cancer recurrence, it is important to note the absence of detailed surgical data. Specific information regarding the type of surgery, completeness of tumor resection, and the use of adjunct therapies, which are critical factors influencing recurrence, is not included in the dataset. This limitation restricts the ability of our predictive models to fully account for the impact of surgical interventions on patient outcomes. Future studies that incorporate detailed surgical variables may offer more robust predictions and should be considered to further enhance the accuracy and applicability of recurrence prediction models in thyroid cancer.

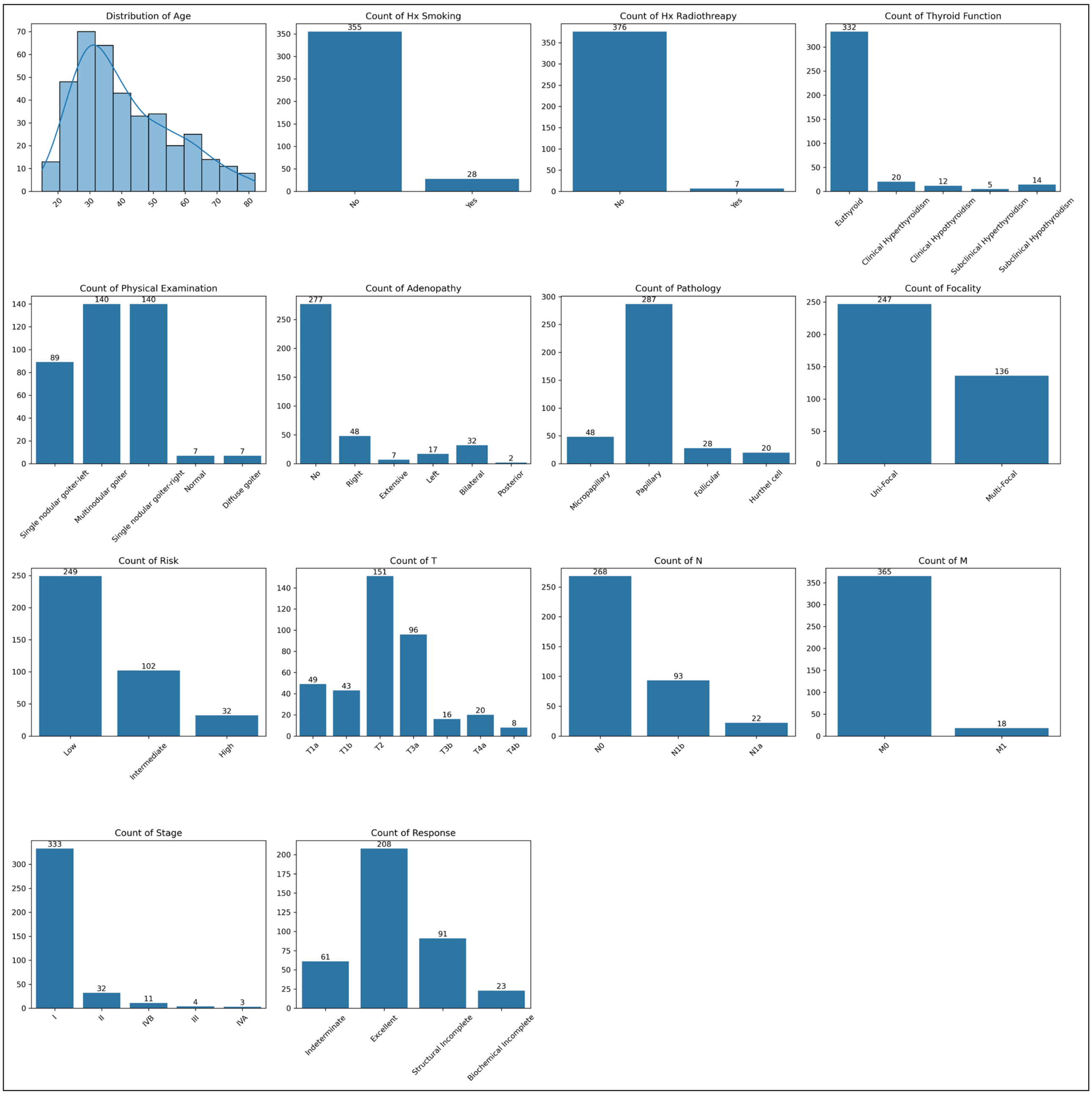

Figure 1. *Dataset features and description.*

Dataset preprocessing

To address the class imbalance in our dataset, we employed the Synthetic Minority Over-sampling Technique (SMOTE). Imbalanced datasets can lead to biased models that favor the majority class, thereby reducing their predictive performance. SMOTE generates synthetic samples for the minority class, effectively balancing the class distribution and ensuring that the model can learn equally from all classes. This technique was crucial for creating a resampled dataset where minority class instances were increased to match the majority class, thus enhancing the model's ability to generalize across different classes [19-25].

Furthermore, we normalized specific numerical features to ensure they are on a comparable scale, which is vital for optimizing model performance. Normalizing the features ensures that they have the same scale, preventing any single feature from disproportionately influencing the model's learning process. This scaling technique prepared our dataset for model training and evaluation, ensuring uniformity and consistency across the selected features. These preprocessing steps significantly contribute to the robustness and reliability of the subsequent machine-learning models [26-28].

Evaluation metrics

To comprehensively evaluate the performance of our classification model, we employed several key metrics: Accuracy, Precision, Recall, F1 Score, receiver operating characteristic (ROC) area under the curve (AUC) Score, and the Confusion Matrix [10,29]. Each of these metrics provides unique insights into the model's effectiveness and reliability.

Accuracy

This metric represents the proportion of correctly predicted instances out of the total instances. It is a straightforward measure of overall correctness but can be misleading in the presence of class imbalance.

Precision

Precision measures the proportion of true positive predictions out of all positive predictions made by the model. It indicates the accuracy of the positive predictions and is particularly important when the cost of false positives is high.

Recall

Also known as sensitivity or true positive rate, recall measures the proportion of true positive predictions out of all actual positives. It is crucial for understanding the model's ability to identify all relevant instances, which is vital in scenarios where missing positive cases is costly.

F1 Score

The F1 Score is the harmonic mean of precision and recall. It provides a balanced measure that considers both false positives and false negatives, making it useful when dealing with imbalanced datasets.

ROC AUC Score

The ROC curve plots the true positive rate against the false positive rate at various threshold settings. The AUC quantifies the overall ability of the model to discriminate between positive and negative classes, with a score closer to 1 indicating better performance.

Confusion matrix

The confusion matrix provides a detailed breakdown of the model's performance by showing the actual versus predicted classifications. It includes true positives, true negatives, false positives, and false negatives, offering a comprehensive view of where the model is making errors. These metrics collectively offer a robust framework for evaluating the classification model, ensuring that both the overall performance and the performance across different classes are thoroughly assessed.

Results and Discussion

To understand the relationships between the various features in our dataset, we generated a correlation heatmap [4] using the Seaborn library. This heatmap visually represents the Pearson correlation coefficients between each pair of features, with values ranging from -1 to 1. A value close to 1 indicates a strong positive correlation, while a value close to -1 indicates a strong negative correlation. Values around 0 suggest no linear correlation between the features.

In the heatmap Figure 2, darker shades of blue indicate strong negative correlations, while darker shades of red represent strong positive correlations. The annotated

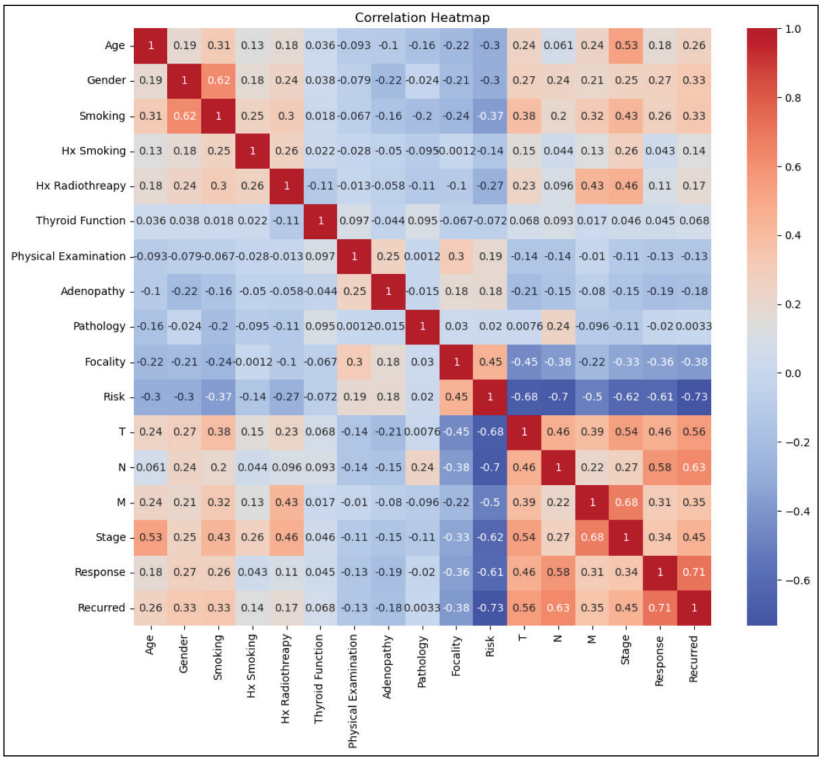

Figure 2. *correlation heatmap between different features.*

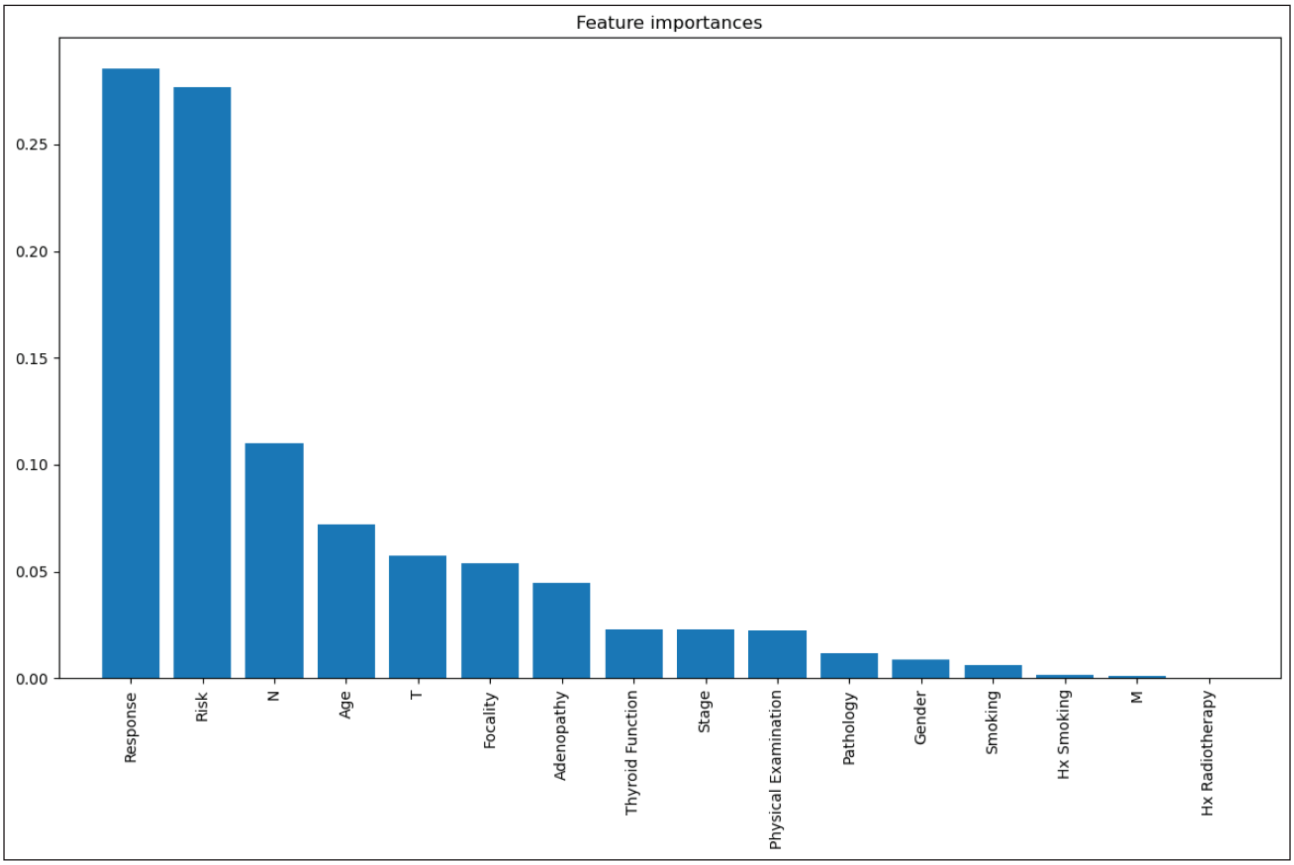

Figure 3. *Feature importance.*

Table 1. *Evaluation matrices for different classification models on test dataset.*

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	0.927,273	0.890,909	0.960,784	0.924,528	0.929,545
K-Nearest Neighbors	0.972,727	0.961,538	0.980,392	0.970,874	0.973,247
Random Forest	0.981,818	0.962,264	1	0.980,769	0.983,051
AdaBoost	0.963,636	0.927,273	1	0.962,264	0.966,102

Abbreviations: ROC, receiver operating characteristic; AUC, Area Under Curve.

values provide the exact correlation coefficients for precise interpretation. This visualization helps identify potential multicollinearity among features and highlights the features most strongly correlated with our target variable, aiding in the feature selection process for model building and interpretation.

Feature importance [7,30] analysis plays a crucial role in understanding the predictive capabilities of ML models. In this study, the feature importance plot (Figure 3) illustrates the relative importance of each feature in predicting the target variable. Features with higher importance values (represented by taller bars) exert greater influence on the model's predictions, indicating their significant role in distinguishing patterns related to the target variable. This visual representation aids in identifying key predictors and understanding their impact on model performance and decision-making processes. Higher feature importance values suggest stronger predictive power and emphasize the relevance of these features in the context of the studied dataset and model architecture.

The performance of various classification models used in this study is summarized in Table 1, which presents the evaluation metrics for each model on the test dataset. The Random Forest model emerged as the best-performing model, achieving the highest accuracy of 0.981818 and a perfect recall score of 1.000000. This indicates that the Random Forest model was highly effective in identifying all instances of thyroid cancer recurrence in the test dataset. Moreover, the precision and F1 Score for the Random Forest model were also notably high, at 0.962264 and 0.980769, respectively, underscoring its balanced performance in both sensitivity and specificity. The ROC AUC score of 0.983051 further validates the superior discriminatory power of the Random Forest model.

In comparison, the KNNs model also showed strong performance, with an accuracy of 0.972727 and a high

F1 Score of 0.970874. AdaBoost achieved commendable results as well, with an accuracy of 0.963636 and a perfect recall score of 1.000000, although its precision and F1 Score were slightly lower than those of the Random Forest model. Logistic Regression, while the least effective among the four, still provided reasonable accuracy and balanced evaluation metrics.

Conclusion

These results highlight the robustness and reliability of the Random Forest classifier in predicting thyroid cancer recurrence, making it a valuable tool for clinical decision-making. The comparative analysis also emphasizes the importance of model selection and evaluation in developing predictive models for medical applications, demonstrating that while all tested models performed well, the Random Forest classifier offered the best overall performance.

Acknowledgment

The author thanks PINUM Cancer Hospital for its support and valuable insights during the research process.

List of abbreviations

Adaboost	Adaptive Boosting
AUC	Area under the curve
CSV	Comma separated values
KNN	K-Nearest Neighbors
ML	Machine learning
ROC	Receiver operating characteristic
SMOTE	Synthetic Minority Over-sampling Technique

Conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Funding

None.

Consent for participate

All data used in this article are publicly available data on the net with no patient identification so consonant for participation is not applicable.

Ethical approval

Ethical approval is not required for articles with publicly available datasets with no patient identification.

Author details

Dilber Iqbal[1], Aamir Shahzad[2]

1. Principal Scientist, Department of Medical Physics, PINUM Cancer Hospital, Faisalabad, Pakistan
2. Modeling and Simulation Laboratory, Department of Physics, Government College University Faisalabad (GCUF), Faisalabad, Pakistan

References

1. Klimaite R, Dauksiene D, Dauksa A, Sarauskas V, Verkauskiene R, Zilaitiene B. The relationship between clinicopathological factors and recurrence risk of papillary thyroid cancer. Endocr Abstr. 2020;70.[Internet]. https://doi.org/10.1530/endoabs.70.EP440
2. Nieto HR, Thornton CE, Brookes K, Nobre de Menezes A, Fletcher A, Alshahrani M, et al. Recurrence of papillary thyroid cancer: a systematic appraisal of risk factors. J Clin Endocrinol Metab. 2022 Apr;107(5):1392–406.https://doi.org/10.1210/clinem/dgab836
3. Rengasamy D, Mase JM, Kumar A, Rothwell B, Torres MT, Alexander MR, et al. Feature importance in machine learning models: a fuzzy information fusion approach. Neurocomputing. 2022;511:511. https://doi.org/10.1016/j.neucom.2022.09.053
4. Kaneko H. Cross-validated permutation feature importance considering correlation between features. Anal Sci Adv. 2022 Sep;3(9-10):278–87. https://doi.org/10.1002/ansa.202200018
5. Oh S. Predictive case-based feature importance and interaction. Inf Sci. 2022;593:155–76. https://doi.org/10.1016/j.ins.2022.02.003
6. Kim SY, Kim Y. Il, Kim HJ, Chang H, Kim SM, Lee YS, et al. New approach of prediction of recurrence in thyroid cancer patients using machine learning. Medicine (United States). 2021;100(42):e27493. https://doi.org/10.1097/MD.0000000000027493
7. Molnar C, König G, Bischl B, Casalicchio G. Model-agnostic feature importance and effects with dependent features: a conditional subgroup approach. Data Min Knowl Discov. 2023;38(5):2903–41. https://doi.org/10.1007/s10618-022-00901-9
8. Giudici P, Gramegna A, Raffinetti E. Machine learning classification model comparison. Socioecon Plann Sci. 2023;87:101560. https://doi.org/10.1016/j.seps.2023.101560
9. An G, Omodaka K, Tsuda S, Shiga Y, Takada N, Kikawa T, et al. Comparison of machine-learning classification models for glaucoma management. J Healthc Eng. 2018;2018:6874765. https://doi.org/10.1155/2018/6874765
10. Baladjay JM, Riva N, Santos LA, Cortez DM, Centeno C, Sison AA. Performance evaluation of random forest algorithm for automating classification of mathematics question items. World J Adv Res Rev. 2023;18(2):034–43. https://doi.org/10.30574/wjarr.2023.18.2.0762
11. Saxena S. Beginners guide to random forest hyperparameter tuning. J Chem Inf Model. 2017;8(9).
12. Nugraha W, Sasongko A. Hyperparameter tuning on classification algorithm with grid search. SISTEMASI. 2022;11(2):391–401. https://doi.org/10.32520/stmsi.v11i2.1750
13. Reddy MN, Latchmana Kumar MR, Kumar PB, Thirumalai S, Nirmala Devi M. Performance enhancement by tuning hyperparameters of random forest classifier for hardware trojan detection. Lect Notes Netw Syst. 2022;288:177–91. https://doi.org/10.1007/978-981-16-5120-5_14
14. Probst P, Wright MN, Boulesteix AL. Hyperparameters and tuning strategies for random forest. Wiley Interdiscip Rev Data Min Knowl Discov. 2019;9(3):e1301. https://doi.org/10.1002/widm.1301
15. Luo R. Improved random forest based on grid search for customer satisfaction prediction. Adv Econ Managt Political Sci. 2023;38(1):198–207. https://doi.org/10.54254/2754-1169/38/20231913
16. Goh RY, Lee LS, Adam MB. Hyperparameters tuning of random forest with harmony search in credit scoring. ASM Sci J. 2019;12:1–9. (Special Issue 5).

17. Dhilsath Fathima M, Samuel SJ. Hyperparameter tuning of ensemble classifiers using grid search and random search for prediction of heart disease. In: Jena OP, Tripathy AR, Elngar AA, Polkowski Z, editors. Computational Intelligence and Healthcare Informatics. Hoboken, NJ:John Wiley & Sons; 2021. pp 139–58. https://doi.org/10.1002/9781119818717.ch8
18. Siji George CG, Sumathi B. Grid search tuning of hyperparameters in random forest classifier for customer feedback sentiment prediction. Int J Adv Comput Sci Appl. 2020;11(9):173–178. https://doi.org/10.14569/IJACSA.2020.0110920
19. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique Nitesh. J Artif Intell Res. 2002;16:321. https://doi.org/10.1613/jair.953
20. Mduma N. Data balancing techniques for predicting student dropout using machine learning. Data (Basel). 2023;8(3):49. https://doi.org/10.3390/data8030049
21. Akın P. A new hybrid approach based on genetic algorithm and support vector machine methods for hyperparameter optimization in synthetic minority over-sampling technique (SMOTE). AIMS Math. 2023;8(4):9400–15. https://doi.org/10.3934/math.2023473
22. Islahulhaq, Wibowo W, Ratih ID. Classification of non-performing financing using logistic regression and synthetic minority over-sampling technique-nominal continuous (SMOTE-NC). Int. J Adv Soft Comput appl. 2021;13(3):115–28. https://doi.org/10.15849/IJASCA.211128.09
23. Elreedy D, Atiya AF, Kamalov F. A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. Mach Learn. 2024;113(7):4903–23. https://doi.org/10.1007/s10994-022-06296-4
24. Xu X, Chen W, Sun Y. Over-sampling algorithm for imbalanced data classification. J Syst Eng Electron. 2019;30(6):1182–91.https://doi.org/10.21629/JSEE.2019.06.12
25. Ogunsanya M, Isichei J, Desai S. Grid search hyperparameter tuning in additive manufacturing processes. Manuf Lett. 2023;35:35. https://doi.org/10.1016/j.mfglet.2023.08.056
26. Shaheen H, Agarwal S, Ranjan P. Minmaxscaler binary pso for feature selection. Adv Intel Syst Comput. 2020:705–16. https://doi.org/10.1007/978-981-15-0029-9_55
27. Nalcin S. StandardScaler versus. minmaxscaler versus. robustscaler: which one to use for your next ML project? Medium. 2022. Available from: https://medium.com/@onersarpnalcin/standardscaler-vs-minmaxscaler-vs-robustscaler-which-one-to-use-for-your-next-ml-project-ae5b44f571b9 . Accessed 2 July 2024.
28. Hale J. Scale, Standardize, or Normalize with Scikit-Learn. Towards Data Science; 2019. https://towardsdatascience.com/scale-standardize-or-normalize-with-scikit-learn-6ccc7d176a02. Accessed 23 Sep 2024.
29. Lever J, Krzywinski M, Altman N. Classification evaluation. Nat Methods. 2016;13(8):603–4.https://doi.org/10.1038/nmeth.3945
30. Saarela M, Jauhiainen S. Comparison of feature importance measures as explanations for classification models. SN Appl Sci. 2021;3(2):272.https://doi.org/10.1007/s42452-021-04148-9