Model Optimization and Tuning Phase Template

Date	25 July 2024
Team ID	SWUID20240034367
Project Title	Anemia-Sense-Leveraging-Machine-Learning-For-Precise-Anemia-Recognitions-using-python
Maximum Marks	10 Marks

Hyperparameter Tuning Documentation (8 Marks):

Model	Tuned Hyperparameters
1. Decision Tree (DT)	• Criterion: The criterion was tuned to 'entropy' instead of the default 'gini' to enhance information gain at each split, leading to a more refined tree. • Max Depth: The maximum depth was carefully adjusted to 10 to prevent the tree from becoming overly complex, which could lead to overfitting. • Min Samples Split: Set to 5 to ensure that each split in the tree was supported by a sufficient number of samples, preventing unnecessary splits. • Min Samples Leaf: Tuned to 2, ensuring that each leaf node had enough samples to represent a valid prediction, thereby reducing overfitting.
2. Random Forest (RF)	• Number of Estimators: Increased to 200 trees to improve the robustness of the model through ensemble learning, allowing for better generalization.

	• Max Features: Set to 'sqrt' to limit the number of features considered for each split, balancing the model's accuracy and training time. • Bootstrap: Enabled to allow sampling with replacement, enhancing the diversity of the individual trees in the forest and reducing variance.
3. Logistic Regression (LR)	• Penalty: The penalty parameter was set to 'l2' (Ridge) to introduce regularization and prevent the model from overfitting to the training data. • Solver: The 'liblinear' solver was chosen for its efficiency with smaller datasets and binary classification tasks. • C (Inverse Regularization Strength): Tuned to 1.0 to balance the trade-off between regularization and model complexity, ensuring the model could capture the essential patterns without overfitting.

Final Model Selection Justification (2 Marks):

Final Model	Reasoning
Random forest	After extensive hyperparameter tuning and model evaluation, the Random Forest (RF) was selected as the final model for anemia detection. • Performance: The Random Forest outperformed other models in terms of accuracy, precision, recall, and F1 score, making it the most reliable model for detecting anemia in the dataset. Its ensemble nature allowed it to effectively capture the complex relationships between the features and the target variable. • Robustness: The model's ability to generalize well across different subsets of the data was demonstrated through cross-validation, with consistently high performance metrics. This robustness was further enhanced by the model's inherent ability to mitigate overfitting through the use of multiple trees. • Interpretability: Although more complex than Logistic Regression, the Random Forest still provided a level of interpretability by allowing us to analyze feature importance, offering insights into which clinical factors were most indicative of anemia. • Computational Efficiency: Despite its complexity, the Random Forest model was computationally efficient and well-suited for deployment, making it a practical choice for real-time anemia detection in a clinical setting.