Team – Ansh Pandey & Shajal

Fine-tunning and finalize

The final phase of the "Anemia Detection with Machine Learning" project was dedicated to optimizing the selected models to enhance their predictive performance and robustness. This phase was crucial as it ensured that the models could generalize well to new data and provide accurate predictions for anemia detection.

Hyperparameter Tuning Documentation

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.

4. K-Nearest Neighbors (KNN)

- **Number of Neighbors**: Set to 5, which was found to provide a balance between bias and variance, offering stable and accurate predictions.

- **Weights**: The 'distance' weight was used, assigning higher importance to closer neighbors, which improved the model's sensitivity to local patterns.
- **Algorithm**: The 'auto' algorithm was selected to automatically determine the best approach for computing nearest neighbors based on the data.

5. Support Vector Machine (SVM)

- **Kernel**: The RBF (Radial Basis Function) kernel was chosen for its ability to handle non-linear relationships in the data, providing a flexible decision boundary.
- **C (Regularization Parameter)**: Set to 1.0 to control the trade-off between maximizing the margin and minimizing classification error, resulting in a balanced model.
- **Gamma**: Tuned to 'scale' to allow the model to adapt to the distribution of the data, improving its ability to detect subtle patterns.

6. Gaussian Naive Bayes (NB)

- **Prior Probabilities**: Adjusted to reflect the distribution of the classes in the training data, ensuring the model could make well-informed predictions even in the presence of class imbalance.
- **Var Smoothing**: Increased to 1e-9 to stabilize the calculation of probabilities, particularly in cases where the variance of a feature was very small, preventing numerical issues.

Final Model Selection Justification

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.

Conclusion

The model was rigorously tested to ensure it met the project's objectives of accurately detecting anemia based on clinical data. The final model was deployed in a live application, allowing healthcare professionals and users to input patient data and receive a prediction on whether the individual is anemic. This deployment marks a significant step toward leveraging machine learning in healthcare, offering a tool that can assist in the early detection and diagnosis of anemia, ultimately leading to better patient outcomes.