MLOps Group No: 92

Group Members Names:

1. PEYALA SAMARASIMHA REDDY - 2023AA05072
2. PEGALLAPATI SAI MAHARSHI - 2023AA05924
3. ANIRUDDHA DILIP MADURWAR - 2023AA05982
4. K VAMSIKRISHNA - 2023AA05209

MLOps Assignment-2

M3: Model Selection & Hyperparameter Optimization - 10M

Justification for the Chosen Model and Hyperparameters:

1. Model Selection:

As we ran for the auto sklearn model, it gave out the best model. The selected model is a **Support Vector Classifier (SVC)** with the Radial Basis Function (RBF) kernel. The SVC is a strong candidate for classification tasks, particularly with high-dimensional data. Its flexibility in finding optimal decision boundaries using kernels makes it suitable for datasets with complex patterns. The model also has a robust mathematical foundation and is known for good generalization performance when hyperparameters are tuned correctly. The best validation accuracy achieved with this model was **0.848**, and the final test accuracy with the tuned hyperparameters is **0.849**, indicating excellent performance and strong generalization capabilities.

2. Hyperparameters and their Significance:

- **C (Regularization Parameter)**: The regularization parameter controls the trade-off between achieving a low training error and a low testing error. A high value of `C` can lead to overfitting, while a low value can result in underfitting. After performing extensive hyperparameter tuning, the optimal value for `C` was found to be **113.11**. This choice strikes a good balance, preventing overfitting while still ensuring a good fit to data.

- **Gamma (Kernel Coefficient)**: Gamma is a parameter that defines how far the influence of a single training example reaches. A small value for gamma can result in a model that is too smooth and underfits, while a large value makes the model more complex, potentially overfitting. The best value for gamma was **0.000117**. This indicates that a

smaller value of gamma was optimal for the dataset, suggesting that the influence of each data point needs to be relatively wide to effectively capture the underlying patterns.

- **Tolerance (tol)**: The tolerance is a stopping criterion for the optimization process. It defines the threshold for the accuracy of the solution when the optimization process will stop. The selected tolerance of **0.00846** ensures that the model runs long enough to converge to a suitable solution, without prematurely stopping or overfitting.

3. Hyperparameter Tuning Process:

The hyperparameter tuning was done using an optimization process involving several trials. The search space for each parameter, particularly `C`, `gamma`, and `tol`, was broad, spanning a range of values to ensure that the best possible combination could be identified. Several trials were conducted, and the model's performance was evaluated using **accuracy** as the primary metric.

- **Trial Performance**: From the various trials, the best-performing set of hyperparameters was found in Trial 19, with **C = 113.11**, **gamma = 0.000117**, and **tol = 0.00846**, yielding the highest validation accuracy of **0.848**. Other trial results varied, with some configurations leading to much lower accuracies. For instance, Trial 3 had a good accuracy of **0.840**, but Trial 2 had a much lower accuracy of **0.099**, highlighting the importance of selecting the right combination of hyperparameters.

4. Ensemble Learning (Optional Consideration):

While the best-performing model was the SVC, the overall performance could potentially be enhanced through ensemble learning methods. These methods combine multiple models to improve performance, often leading to better generalization and reduced overfitting. However, for this analysis, the SVC with the tuned hyperparameters provided the best accuracy.

5. Final Model Performance:

The final chosen model and its hyperparameters resulted in a test accuracy of **0.849**, which is consistent with the best validation accuracy. This performance demonstrates that the model has not only learned well from the training data but has also generalized effectively to unseen data, making it a robust classifier.

6. Conclusion:

In summary, the selected **Support Vector Classifier (SVC)** with the **RBF kernel** and the hyperparameters `C = 113.11`, `gamma = 0.000117`, and `tol = 0.00846` offers an optimal solution for this classification task. The model is highly effective in classifying the data, achieving **0.849** accuracy on the test set, with strong generalization capabilities. The successful tuning of the hyperparameters, especially the selection of a moderately high `C` and a small `gamma`, helped the model strike a balance between bias and variance, leading to high performance across training and test sets.