Model Validation, Overfitting Control & Hyperparameter Tuning

Introduction

This report documents the implementation and evaluation of Task–3, focusing on model validation techniques, overfitting detection, and hyperparameter tuning using regression models on the California Housing dataset and baseline models from Task–2 are used to ensure fair and scientific comparison.

Methodology

The California Housing dataset is used to maintain continuity with Task–2. The dataset is split into training and testing subsets. Feature scaling is performed using Standard Scaler to normalize the input features.

A baseline Linear Regression model is trained first. Cross-validation with 5 folds is applied to validate model performance. Overfitting is analyzed using a Decision Tree Regressor by comparing training and testing scores. Grid Search CV is used to tune the hyperparameter of Ridge Regression.

Overfitting Analysis

Overfitting is analyzed using a Decision Tree Regressor. The model achieves very high training R^2 scores but significantly lower testing R^2 scores. This gap clearly indicates overfitting, where the model memorizes training data but fails to generalize well to unseen data.

This analysis highlights why relying only on training accuracy can be misleading and why validation techniques are essential in real-world machine learning systems.

Hyperparameter Tuning Approach

To control overfitting and improve generalization, hyperparameter tuning is applied to Ridge Regression, an extension of Linear Regression from Task–2. Grid Search CV is used to systematically test multiple values of the regularization parameter alpha using cross-validation.

Cross-validation ensures that each model configuration is evaluated on multiple data splits, providing a reliable estimate of performance. The optimal alpha value balances bias and variance, leading to improved stability.

Results

Cross-validation results demonstrate consistent R^2 scores, confirming model reliability. Decision Tree Regression shows high training accuracy but lower testing accuracy, indicating overfitting. Hyperparameter tuning improves Ridge Regression performance.

The optimized Ridge Regression model provides better generalization performance than the baseline Linear Regression model.

Final Model Selection & Justification

The optimized Ridge Regression model is selected as the final model. Compared to baseline Linear Regression from Task–2, the tuned Ridge model achieves improved R^2 scores and lower error on unseen data.

Unlike Decision Tree Regression, which overfits the training data, Ridge Regression provides a balanced solution with better generalization. Therefore, the tuned Ridge model is considered more reliable and production-ready for house price prediction.

Conclusion

This task demonstrates the importance of validation, overfitting control, and hyperparameter tuning in machine learning. Building upon Task–2, Task–3 ensures that the selected model is not only accurate but also reliable. The final optimized Ridge Regression model satisfies industry-level expectations for robustness and generalization.