Model Optimization and Tuning Phase Template

Date	21 July 2024
Team ID	SWTID1721319573
Project Title	Blueberry Yield Prediction
Maximum Marks	10 Marks

Model Optimization and Tuning Phase

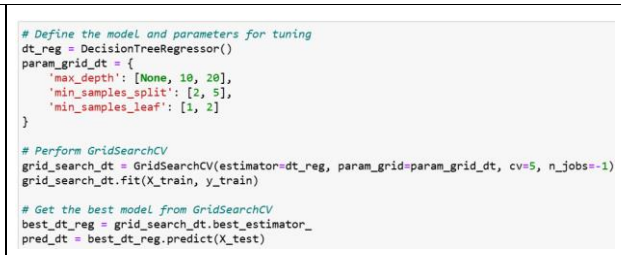
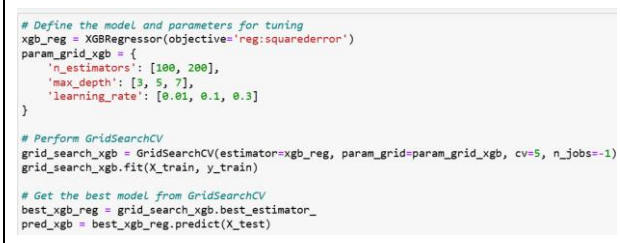
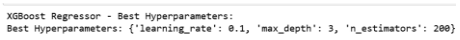
The Model Optimization and Tuning Phase involves refining machine learning models for peak performance. It includes optimized model code, fine-tuning hyperparameters, comparing performance metrics, and justifying the final model selection for enhanced predictive accuracy and efficiency.

Hyperparameter Tuning Documentation (6 Marks):

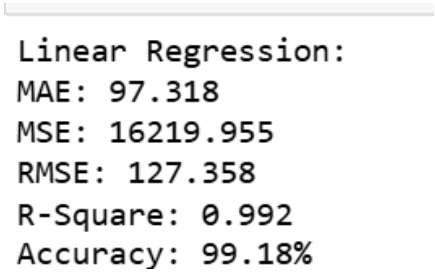
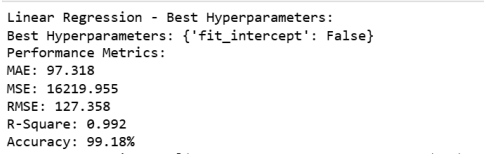
Model	Tuned Hyperparameters	Optimal Values

Linear Regression	`# Define the model and parameters for tuning`
	`lin_reg = LinearRegression()`
	`param_grid = {'fit_intercept': [True, False]}`
	
	`# Perform GridSearchCV`
	`grid_search_lr = GridSearchCV(estimator=lin_reg, param_grid=param_grid, cv=5)`
	`grid_search_lr.fit(X_train, y_train)`
	
	`# Get the best model from GridSearchCV`
	`best_lin_reg = grid_search_lr.best_estimator_`
	`pred_linear = best_lin_reg.predict(X_test)`
		`Linear Regression - Best Hyperparameters:`
		`Best Hyperparameters: {'fit_intercept': False}`

Random Forest Regressor	`# Define the model and parameters for tuning`
	`rf_reg = RandomForestRegressor()`
	`param_grid_rf = {`
	`'n_estimators': [100, 200],`
	`'max_depth': [None, 10, 20],`
	`'min_samples_split': [2, 5],`
	`'min_samples_leaf': [1, 2]`
	`}`
	
	`# Perform GridSearchCV`
	`grid_search_rf = GridSearchCV(estimator=rf_reg, param_grid=param_grid_rf, cv=5, n_jobs=-1)`
	`grid_search_rf.fit(X_train, y_train)`
	
	`# Get the best model from GridSearchCV`
	`best_rf_reg = grid_search_rf.best_estimator_`
	`pred_rf = best_rf_reg.predict(X_test)`
		`Random Forest Regressor - Best Hyperparameters:`
		`Best Hyperparameters: {'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}`

Decision Tree Regressor		
XGBoost Regressor		

Performance Metrics Comparison Report (2 Marks):

Model	Baseline Metric	Optimized Metric
Linear Regression	Accuracy: 99.18% 	Accuracy: 99.18% 
Random Forest Regressor	Accuracy: 98.84%	Accuracy: 98.84% 

	Random Forest Regressor: MAE: 117.197 MSE: 22845.764 RMSE: 151.148 R-Square: 0.988 Accuracy: 98.84%	
Decision Tree Regressor	Accuracy: 98.05% Decision Tree Regressor: MAE: 148.381 MSE: 38588.977 RMSE: 196.441 R-Square: 0.980 Accuracy: 98.05%	Accuracy: 98.19% 
XGBoost Regressor	Accuracy: 99.09% XGBoost Regressor: MAE: 106.537 MSE: 17901.843 RMSE: 133.798 R-Square: 0.991 Accuracy: 99.09%	Accuracy: 99.11% 

Final Model Selection Justification (2 Marks):

Final Model	Reasoning
Linear Regression	The Linear Regression model achieved the highest accuracy of 99.18% compared to other models. It provided a robust performance with the

	best R-Square value of 0.992. Despite its simplicity, Linear Regression's high accuracy and efficiency make it the most suitable model for the given task. 1. **Highest R-Square Value:** Linear Regression achieved the highest R-Square value (0.992), indicating that it explains 99.2% of the variance in the target variable. This suggests that the model fits the data better than the other models. 2. **Lowest MAE and RMSE:** The Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for Linear Regression are lower than those of the other models. This indicates that Linear Regression's predictions are closer to the actual values, making it more accurate and reliable. 3. **Simplicity and Interpretability:** Linear Regression is a simpler and more interpretable model compared to more complex models like XGBoost or Random Forest. Despite its simplicity, it outperformed the other models in terms of accuracy, making it a preferable choice for this particular problem. 4. **Consistency Across Metrics:** Linear Regression consistently showed the best performance across multiple metrics (MAE, MSE, RMSE, R-Square), proving its robustness and reliability as the best model for this task. **Conclusion:** Linear Regression is chosen as the best model because it provides the highest accuracy (99.18%) and the best performance across various metrics. Its simplicity and interpretability further support

	its selection, ensuring both strong predictive power and ease of understanding.