Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

Machine Problem No. 3			
Topic:	**Evaluating Machine Learning Model Performance**	**Week No.**	
Course Code:	**CSST102**	**Term:**	1st Semester
Course Title:	**Basic Machine Learning**	**Academic Year:**	2025-2026
Student Name	Capili, Judeelyn M	**Section**	BSCS 3A
Due date		**Points**	

Evaluating Machine Learning Model Performance Using Logistic Regression

In this activity, I used the California Housing dataset from Scikit-Learn to evaluate the performance of a Logistic Regression model. Although the dataset is originally designed for regression, it was transformed into a binary classification problem to meet the requirements of the task. Each data point in the dataset represents a housing district in California and includes eight key features: median income, house age, average number of rooms, average number of bedrooms, population, average occupancy, latitude, and longitude. The target variable was created by labeling houses as “High Value” (1) if their median house value was above the dataset’s median, and “Low Value” (0) otherwise.

Before training, the data was preprocessed to ensure high quality and consistency. Missing values were handled appropriately, and all numerical features were standardized using Scikit-Learn’s StandardScaler to normalize their ranges. This preprocessing step ensured that no single variable dominated the learning process due to scale differences. The dataset was then split into 80% training data and 20% testing data, ensuring that both subsets were representative of the entire dataset distribution.

The Logistic Regression model was then trained on the standardized training data. The model achieved a training accuracy of 0.842 and a testing accuracy of 0.832, showing that it learned the underlying patterns effectively and maintained consistent performance on unseen data. This small difference between training and testing accuracy suggests that the model generalized well with minimal overfitting.

Cross-Validation (5-Fold) Results

To validate the model's reliability and ensure its performance was not dependent on a single train-test split, I performed a 5-Fold Cross Validation. This approach divides the dataset into five equal parts, trains the model on four parts, and tests it on the remaining part, repeating the process five times.

The cross-validation results showed a mean accuracy of 0.832 with a standard deviation of 0.011. These results demonstrate that the model is stable and performs consistently across different subsets of data. A low standard deviation indicates that the model's accuracy does not fluctuate significantly between folds, meaning it can be trusted to perform similarly on new, unseen data. This consistency is crucial in real-world applications where models must make reliable predictions under varied conditions.

Confusion Matrix and Metrics

The confusion matrix visualization revealed that the Logistic Regression model correctly classified most "High Value" and "Low Value" houses, with only a small number of misclassifications. It highlights how the model performed in distinguishing between the two classes, and it provides insights beyond simple accuracy.

The model's overall accuracy was 0.83, while its precision was 0.82, recall was 0.84, and F1-score was 0.83. These results indicate that the model balanced false positives and false negatives effectively. Precision shows that 82% of houses predicted as "High Value" were actually high-value houses, while recall means that the model identified 84% of all truly high-value homes. The F1-score, which combines precision and recall into a single metric, confirms the model's balanced performance.

In practical terms, the confusion matrix shows that the model rarely misclassifies high-value houses as low-value ones or vice versa. This is important because, in real-world housing analysis, misclassifying property values can lead to poor investment decisions or inaccurate policy planning. The near-equal balance between precision and recall also implies that the model does not favor one class over another, which is ideal for fair and unbiased predictions.

Learning Curve Interpretation

The learning curve visualization shows the relationship between model accuracy and the amount of training data used. The curve revealed that as the number of training samples increased, both the training and cross-validation scores gradually converged and stabilized around 0.83 accuracy. This indicates that the model has learned effectively from the data and reached an optimal performance point where additional data yields diminishing returns in accuracy.

The convergence of the two curves (training and validation) suggests that the model is neither overfitting nor underfitting. Overfitting occurs when a model performs well on training data but poorly on testing data, while underfitting happens when the model is too simple and fails to capture important data patterns. In this case, the close alignment of the curves shows that the Logistic Regression model has a good balance between bias and variance. It generalizes well to unseen data and performs consistently as more training samples are added.

This pattern also suggests that the chosen model and preprocessing techniques are appropriate for the dataset. If the learning curve had shown a large gap between training and validation performance, it would have indicated the need for more regularization or data cleaning. Conversely, if both lines were low and flat, it would suggest that the model lacks complexity and cannot learn from the data effectively.

Model Comparison Visualization

The comparison chart (comparison.png) presents the differences between the model's training accuracy, testing accuracy, and cross-validation mean accuracy. The three bars are nearly identical in height, indicating that the model performs consistently across all evaluation scenarios.

This comparison confirms that the Logistic Regression model did not overfit to the training data and that its generalization to unseen data is strong. When a model's training accuracy is much higher than its testing accuracy, it usually indicates overfitting; however, in this case, both values are very close, showing stable learning behavior. The inclusion of the cross-validation score adds further confidence that the model's strong performance is not coincidental but repeatable across different data splits.

In real-world terms, this consistency implies that the model could reliably classify housing values in different California districts, even those not included in the training data. Such reliability is essential for decision-making tasks like housing market predictions, policy assessments, or property investment evaluations.

Interpretation and Discussion

What do the results of the confusion matrix indicate?

The results of the confusion matrix indicate that the Logistic Regression model is highly effective at distinguishing between "High Value" and "Low Value" houses. It correctly classified the majority of cases, with very few false positives or false negatives. This means that the model has a well-defined decision boundary and can make accurate predictions for both categories. The balanced precision and recall scores show that the model treats both classes fairly, which is important in preventing bias. In a real-world context, this means the model could be used reliably by real estate analysts or policymakers to assess housing market trends with minimal misclassification errors.

How consistent is the model's performance based on 5-Fold Cross Validation?

The 5-Fold Cross Validation results confirm the model's high level of consistency and stability. The mean accuracy of 0.832 and a standard deviation of 0.011 demonstrate that the model performs almost equally well across different subsets of data. This means the model's performance is not overly dependent on specific samples from the training set, which is a sign of a well-generalized model. Such stability is critical in professional applications where data variability can affect performance; for example, in housing market analysis, the model's consistency ensures reliable predictions across different regions or time periods.

What insights can be derived from the learning curve?

The learning curve shows that the model's performance improves as it sees more training data but eventually stabilizes, indicating that it has reached its optimal learning capacity. The close convergence between the training and validation lines signifies a healthy balance between bias and variance. This means the model has learned the essential relationships between the input features and the target variable without memorizing the data. In practical terms, this suggests that Logistic Regression is well-suited for this dataset's complexity and that collecting significantly more data would not drastically improve performance. The learning curve thus provides evidence of a well-trained and well-calibrated model.

How can the model be improved?

While the Logistic Regression model performs well, there are several ways to improve it further. First, more advanced algorithms such as Random Forests, Support Vector Machines (SVM), or Gradient Boosting could capture non-linear relationships that Logistic Regression may miss. Second, performing hyperparameter tuning through Grid Search or Random Search could optimize model parameters, improving accuracy and reducing variance. Third, additional feature engineering could be applied, for example, adding interaction terms or transforming skewed variables to help the model capture deeper patterns in the data. Lastly, techniques like regularization (L1 or L2) or dimensionality reduction (PCA) could help control model complexity and prevent even minor overfitting.

Short Reflection:

In this activity, I successfully applied Logistic Regression to a real-world dataset and evaluated its performance using cross-validation, confusion matrix analysis, and learning curve visualization. The model achieved consistent results across all metrics, showing strong generalization and minimal overfitting. I learned how performance evaluation techniques provide deeper insights into a model's strengths and weaknesses beyond simple accuracy. This exercise helped me understand how to interpret key metrics like precision, recall, and F1-score, and how they relate to practical decision-making. Overall, this project deepened my appreciation for data preprocessing, model evaluation, and the importance of validation in building reliable machine learning systems.