Weekly Report #4

This week, we worked with two datasets and four machine learning (ML) models: Logistic Regression, Support Vector Machine (SVM), Random Forest, and Decision Tree. Each dataset was used to implement a separate model for each of these ML techniques. We applied validation techniques like cross-validation and learning curves to evaluate the models' generalization and performance.

Logistic Regression Model

- Dataset 1 (1000 x 14): The model performed well with a high accuracy of 0.96. The cross-validation score was also 0.96, and the learning curve indicated a good fit with slight bias.
- Dataset 2 (1888 x 14): The model showed lower accuracy (0.76 on training data and 0.71 on test data). The cross-validation score was 0.77, and the learning curve suggested underfitting with high bias.
- Optimization Attempts: We applied data cleanup techniques like the interquartile range (IQR) to remove outliers, used GridSearch for hyperparameter tuning, and added L2 (ridge) regularization to prevent overfitting. Despite these efforts, the model remained underfitted.

Decision Tree Model

- Dataset 1 (1000 x 14): The model achieved 0.94 accuracy, with a cross-validation score of 0.95. The learning curve indicated a good fit with slight bias.
- Dataset 2 (1888 x 14): The model performed very well, achieving 1.00 accuracy on training data and 0.94 on test data. The cross-validation score was 0.94, and the learning curve confirmed a good fit.
- Cross-Dataset Testing: When we trained the model on Dataset 2 and tested on Dataset 1, the accuracy dropped to 0.69, although the learning curve still showed a good fit.

Support Vector Machine (SVM) Model

- Dataset 1 (1000 x 14): The model performed well, achieving 0.96 accuracy with a cross-validation score of 0.96. The learning curve showed a good fit with slight bias.

- Dataset 2 (1888 x 14): The model showed high accuracy (0.90 on training data and 0.87 on test data). The cross-validation score was 0.86, and the learning curve indicated a good fit.

Random Forest Model

- Dataset 1 (1000 x 14): This model performed the best, achieving 0.98 accuracy with a cross-validation score of 0.97. The learning curve indicated a good fit with slight bias.
- Dataset 2 (1888 x 14): The model performed well on training data but had a lower accuracy of 0.74 on test data. The cross-validation score was 0.92, and the learning curve indicated a good fit.

Summary

- Best Overall Model: The Decision Tree and Random Forest models showed the highest performance across both datasets.
- Cross-Dataset Testing Issue: Training on Dataset 2 and testing on Dataset 1 resulted in significantly lower accuracy, indicating possible dataset-specific patterns.
- Underfitting Issue in Logistic Regression: Despite optimization efforts, the Logistic Regression model remained underfitted for Dataset 2.