Independent Model Development: Logistic Regression

Dataset:

This study uses the *Gas Sensor Array under Low Concentration (gslac)* dataset from the UCI Machine Learning Repository. The dataset consists of responses from 10 metal oxide semiconductor sensors exposed to six gases (ethanol, acetone, toluene, ethyl acetate, isopropanol, and n-hexane) at three concentration levels (50, 100, and 200 ppb). Each sample contains 9000 sensor response points, concatenated across sensors.

Objective:

The goal was to perform an initial model exploration of the *gslac* data using different machine learning approaches. Logistic Regression and Neural Networks were selected as baseline models. This work builds on the data readiness and exploratory analyses completed in previous sprints.

Model Rationale:

Logistic Regression models the posterior probability of class membership as:

$$p(C_k \mid x) = \exp(w_k^T x) / \Sigma_j \exp(w_j^T x)$$

where each class k has a weight vector w_k.
The model learns these weights by maximizing the conditional likelihood, effectively finding linear decision boundaries that best separate the classes in feature space.It provides probabilistic predictions and interpretable boundaries, making it a strong baseline before exploring more flexible models such as Neural Networks.

Methodology:

Using scikit-learn's *LogisticRegression* with the L-BFGS solver and a multinomial objective, the model was trained on 80% of the samples (using a stratified split) and evaluated on the remaining 20%. Sensor features were standardized with *StandardScaler* to improve numerical stability and convergence. Model hyperparameters were tuned experimentally, including *max_iter* and the regularization strength *C*.

Model Training and Evaluation:

The multinomial logistic regression model achieved 100% classification accuracy across all six gases, with perfect alignment between predicted and true labels. The confusion matrix showed no off-diagonal errors, and the classification report reported precision, recall, and F1-scores of 1.00 for each gas type. This suggests that the sensor response patterns are highly distinctive and linearly separable. Although perfect accuracy can sometimes indicate overfitting, 3-fold cross-validation yielded similarly perfect results, confirming that the model generalized well. Regularization tuning (via *C*) and solver variations (e.g., *saga*, *newton-cg*) showed minimal impact, reinforcing the robustness of the logistic regression model.

Comparison with Neural Network Model:

A parallel experiment by a team member used a feed-forward neural network built with TensorFlow/Keras. Despite varying the number of hidden layers (2–3) and activation functions (*sigmoid*, *ReLU*), the network achieved slightly lower accuracy, around 0.88. This suggests that, for this dataset, the linear decision boundaries of logistic regression are sufficient to capture the discriminative relationships between sensor responses.

Future Work

Next steps include investigating regression-based prediction of gas concentrations rather than discrete classification, evaluating performance under sensor noise or drift.