Part A: Linear Regression

Design and Approach

The objective for Part A was to model the relationship between independent variables and a continuous target. The approach involved minimizing the Mean Squared Error (MSE) to find the optimal weights for the linear model:

MSE = (1/n) \sum_{i=1}^n (y_i - (w x_i + b))^2

The solution was derived either through Gradient Descent, iteratively updating parameters, or the Normal Equation for a direct analytical solution.

Model and Parameters

- **Model:** Linear Regression ($y = w^T x + b$).
- **Parameters:** Learning rate (alpha) and number of iterations (if using Gradient Descent), or the feature weights (w) and bias (b).
- **Evaluation:** Performance was measured using Mean Squared Error (MSE) and R-squared (R^2) scores to determine the goodness of fit.

Part B: k-Nearest Neighbors (Lenses and Credit Approval)

Design and Approach

The design for Part B focused on a classification algorithm that relies on the local structure of the data.

- **Preprocessing:** Significant effort was placed on the Credit Approval dataset, using **label-conditioned mean imputation** for missing numerical values and **Z-score normalization** to prevent features with large ranges from dominating the distance metric.
- **Hybrid Distance:** A custom L2 distance was implemented to handle mixed data types:
    - **Numerical:** Squared difference $(a_i - b_i)^2$.
    - **Categorical:** Binary distance (1 if different, 0 if the same).

Evaluation Results (Accuracy)

Dataset	**k=1**	**k=3**	**k=5**	**k=7**
Lenses	0.8000	0.7500	0.7500	0.6250

Credit Approval	0.8116	0.8478	0.8333	0.8478

Part C: Spam Identification (Naive Bayes)

Design and Approach

For the spam detection task, a Gaussian Naive Bayes classifier was designed. The approach assumed that each feature (word or character frequency) follows a normal distribution within each class (Spam vs. Non-Spam).

- **Log-Space Arithmetic:** To avoid numerical underflow during the multiplication of multiple small probabilities, the model calculated the **log-likelihood** plus the **log-prior**.

Model and Parameters

- **Model:** Gaussian Naive Bayes.
- **Parameters:** Class-specific means (mu) and variances (sigma squared) for each of the 57 features in the Spambase dataset.
- **Evaluation:** In addition to accuracy, Precision and Recall were prioritized to assess the cost of False Positives (legitimate email incorrectly classified as spam).

Metric	**Value**
Accuracy	0.8241
Precision	0.7105
Recall	0.9082
F1-Score	0.7972

Reflection

The completion of these three parts revealed the specific strengths and weaknesses of different modeling strategies. Linear Regression is highly interpretable but limited by its assumption of linearity. k-Nearest Neighbors is flexible and intuitive but highly dependent on proper feature

scaling and the choice of k, as seen in the sensitivity of the Credit Approval results. Naive Bayes is remarkably fast and effective for high-dimensional data like spam detection, even though the "naive" independence assumption is rarely perfectly met in real-world text.

The most critical insight was the role of preprocessing. The difference between a raw dataset and one that has been correctly imputed and normalized was the primary factor in achieving high accuracy, particularly in Parts B and C.