Use Case: Bank Check Processing Automation

Business Problem:
In the banking industry, processing checks remains a critical task that is predominantly manual and labor-intensive. Each day, bank employees are required to read and input handwritten amounts from thousands of checks into the bank's systems. This process is prone to human errors, such as misreading digits or entering incorrect amounts, leading to financial discrepancies and customer dissatisfaction. Furthermore, the manual nature of this process results in high labor costs and inefficiencies, reducing the overall productivity of bank operations.

Objective:
The primary objective is to develop an automated system using machine learning models to recognize and digitize handwritten amounts on bank checks. By leveraging the MNIST dataset, which contains images of handwritten digits, the goal is to build a robust classifier that can accurately interpret and convert these handwritten amounts into machine-readable text. This solution will reduce errors, enhance operational efficiency, and lower the dependency on manual labor in the check processing workflow.

Project Summary:

Data Understanding:
The MNIST dataset, consisting of 70,000 grayscale images of handwritten digits (0-9), is used as the basis for training the model. Each image is 28x28 pixels, and the dataset is well-balanced, with roughly the same number of examples for each digit class. This balance ensures that the model will be trained on an evenly distributed dataset, reducing the risk of bias towards any particular digit.

Data Preparation:
The dataset was normalized by scaling the pixel values to a range of [0, 1], ensuring that all features contribute equally to the model's learning process. The data was then split into training and testing sets to evaluate the model's performance. Additionally, checks for missing values and duplicates were conducted, though the MNIST dataset is generally clean and does not contain such issues.

Modeling:
Three models were developed to recognize and digitize handwritten amounts on bank checks:
1. Gaussian Naive Bayes (GaussNB): This model assumes that features are independent and normally distributed within each class. It is simple, computationally efficient, and suitable for high-dimensional datasets like MNIST.

2. Gaussian Bayes with Full Covariance (GaussBayes): This model accounts for correlations between features by using the full covariance matrix, allowing the model to capture more complex relationships in the data.
3. K-Nearest Neighbors (KNN): A non-parametric model that classifies digits based on the closest training examples in the feature space. This model is particularly effective for image recognition tasks due to its ability to capture local structure in the data.

Evaluation:
The models were evaluated primarily using accuracy as the metric to determine how well the models were able to classify the digits correctly. Accuracy is defined as the ratio of correctly predicted instances to the total instances in the dataset. The results showed that KNN outperformed the others models with 96% of accuracy, making it the most effective model for this use case.

Conclusion:
The implementation of this automated check processing system represents a significant step towards modernizing bank operations. By reducing the reliance on manual data entry, the system not only improves accuracy and efficiency but also allows bank employees to focus on more complex and value-added tasks. The use of the MNIST dataset ensures that the model generalizes well to unseen data, making it a reliable tool for real-world deployment in check processing workflows. The KNN model, in particular, provided the highest accuracy, indicating its suitability for this digit recognition task.