Summary

The objective of this project was to create a machine learning model that automatically recognizes handwritten digits in order to expedite the data entry process for customer surveys. In situations where surveys are often completed on paper, the main goals were to decrease human labour, minimise errors, and speed up data processing.

I developed and trained the model using the MNIST dataset, which is made up of pictures of handwritten numbers. The subsequent actions were carried out:

1. **Data Preparation**:
    - The dataset was preprocessed using **Pandas** for data manipulation and **NumPy** for efficient numerical operations.
    - Normalised the pixel values of the images to a range between 0 and 1, ensuring that the model trained more effectively and converged faster.
    - To address class imbalance, I applied an undersampling technique, where each class was sampled to match the size of the smallest class, ensuring a balanced representation of all digits during training.
2. **Modeling**:
    - I implemented a **K-Nearest Neighbors (KNN)** classifier with 5 neighbours, which is a simple yet effective algorithm for image recognition tasks.
    - The model was trained to recognize and classify handwritten digits, simulating the task of identifying numbers from consumer survey responses.
3. **Evaluation**:
    - The model's performance was evaluated using the accuracy metric, achieving an impressive accuracy of **96.9%** on the test set.
4. **Tools Used**:
    - **Pandas** and **NumPy** were pivotal in data handling, preprocessing, and ensuring efficient computation throughout the project.
    - The **KNN** algorithm was implemented for classification, and the model's effectiveness was validated using accuracy as the primary evaluation metric.

This project successfully demonstrated the feasibility of using machine learning to automate and enhance the efficiency of consumer survey data entry processes, offering significant improvements over traditional manual methods.