Saman Aijaz Siddiqui (B21BB027)

Low Precision Machine Learning

Introduction

It is an in-depth analysis of a machine learning model implementation where input data is quantized to lower precision levels. The objective is to understand the impact of quantization on model performance, including variations in data representation across 8-bit, 4-bit, and 2-bit precision.

This approach is tested on the Iris dataset, which is widely used for classification tasks. The report discusses quantization techniques, evaluates the effect on model accuracy, and presents visualizations for performance at different precision levels.

Objective and Scope

The goal is to:

1. Implement data quantization at various bit precisions.
2. Train and evaluate a classification model on quantized data.
3. Assess the model's accuracy and computational efficiency under different quantization settings.

Data and Preprocessing

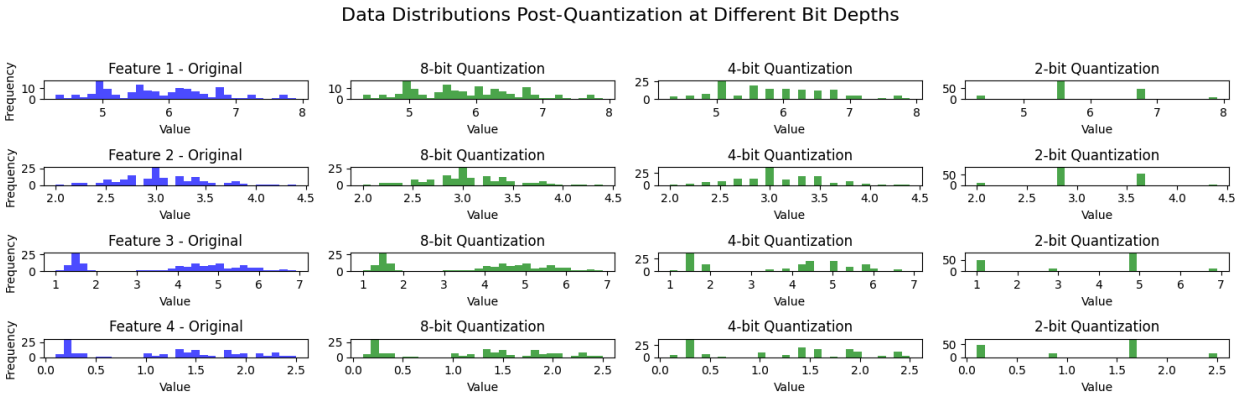
Dataset Description

- **Dataset Used**: Iris dataset
- **Features**: 4 numerical features representing various flower measurements.
- **Target Classes**: Three classes, each corresponding to a unique Iris flower species.

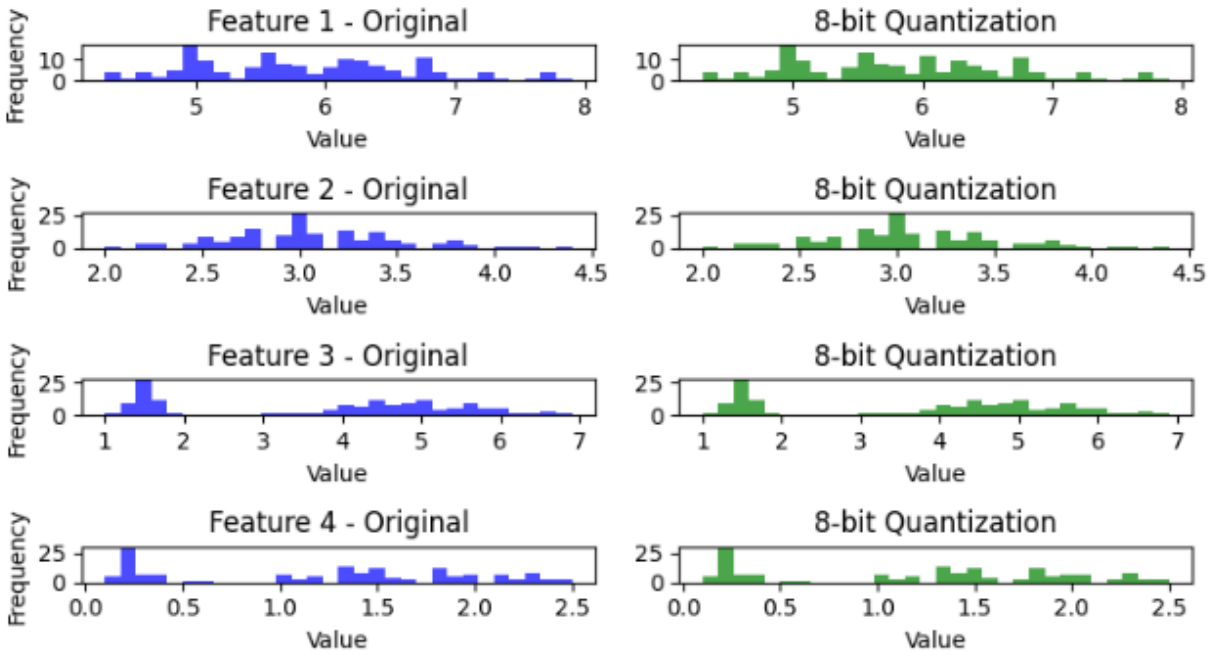
Data Quantization Process

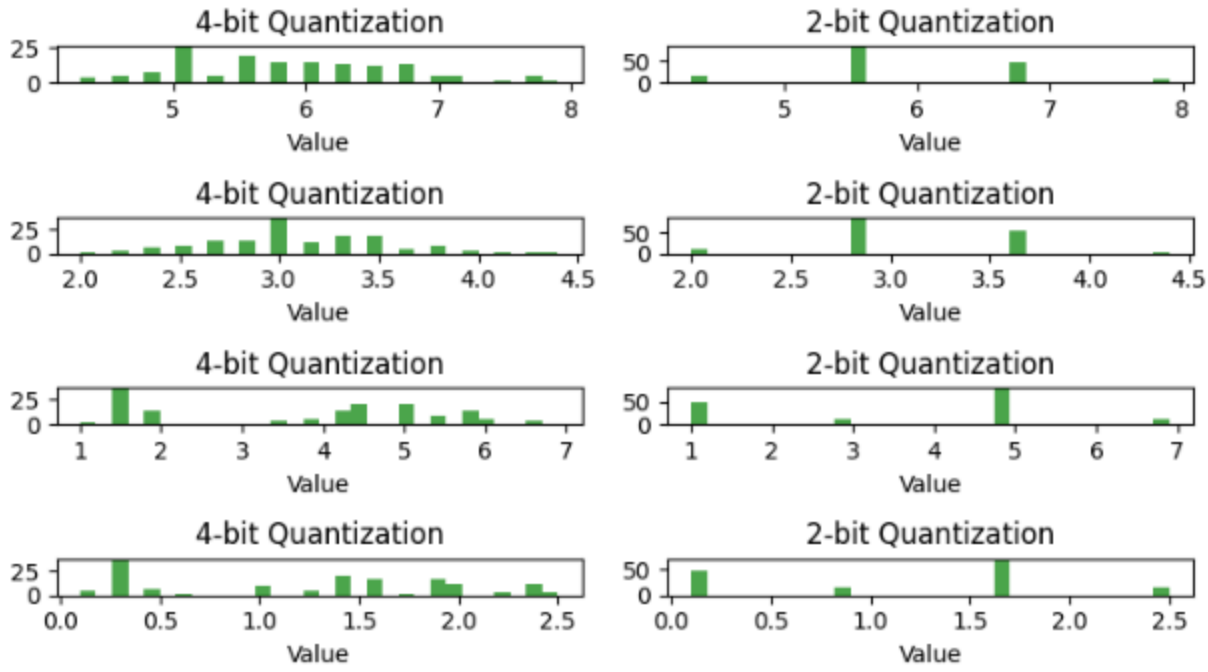
The data quantization process converts the continuous feature values into discrete levels based on a specified bit depth:

- **8-bit Quantization**: Allows for higher precision, with 2^8 = 256 discrete levels.
- **4-bit Quantization**: Moderate precision, reducing levels to 2^4 = 16.
- **2-bit Quantization**: Low precision, with only 2^2 = 4 levels.



Zoomed in images:





Model Training and Evaluation

Model Selection

A decision tree classifier is employed to evaluate the impact of data quantization on model accuracy. This model is suitable for interpretability and works well with discrete, quantized data.

Cross-Validation

The model undergoes cross-validation, where accuracy is evaluated across multiple splits to ensure robustness. This approach provides insights into the average performance under each quantization level.

Results and Analysis

Performance Metrics

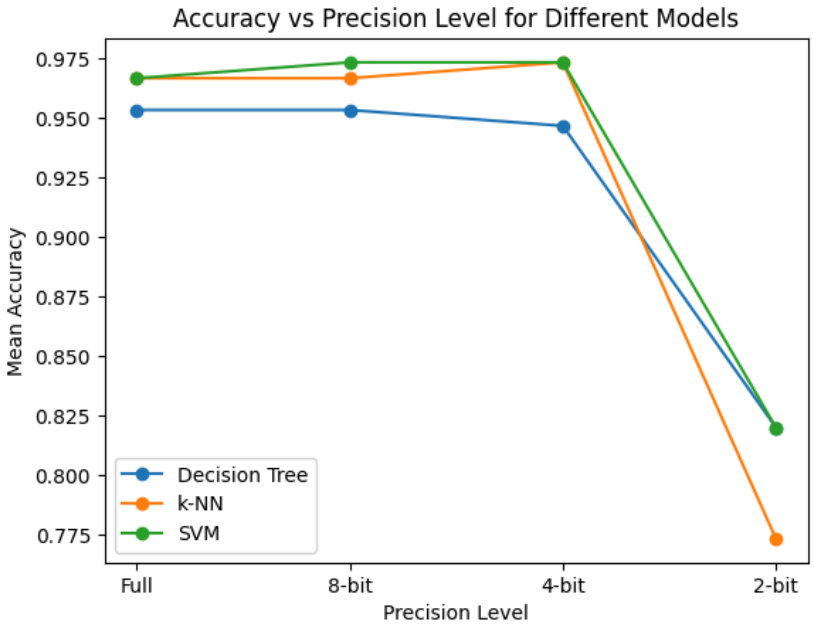
The quantized data's impact on model accuracy is evaluated and compared across precision levels. Metrics reported include:

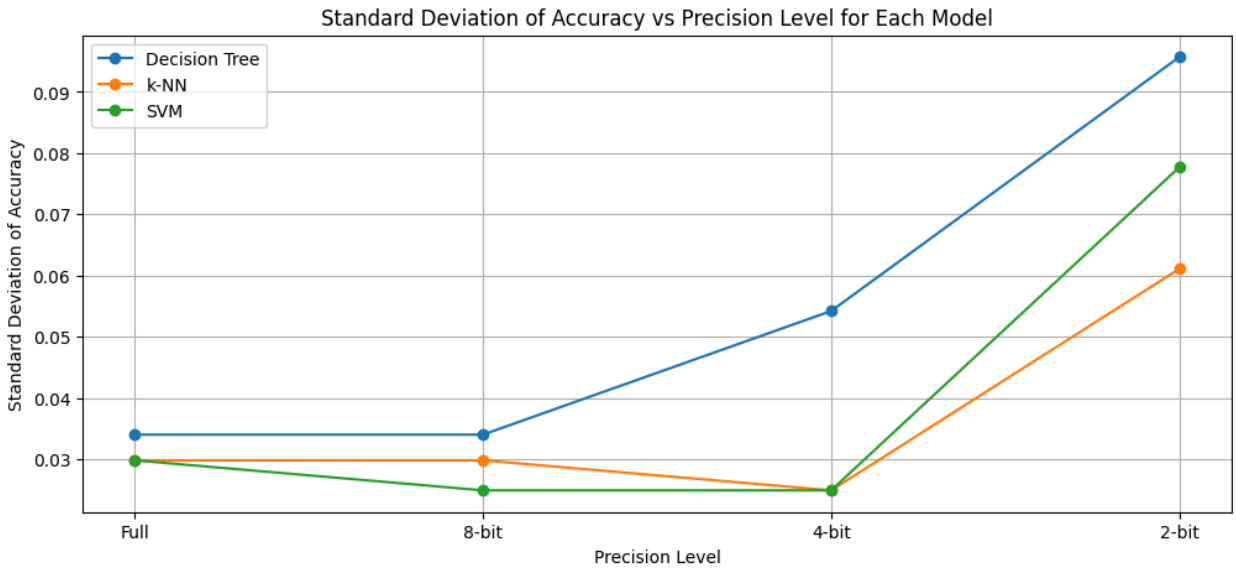
- **Classification Accuracy**: For each bit depth.
- **Computational Efficiency**: Time and memory use reduction.

```
Model: Decision Tree
Full Precision - Mean Accuracy: 0.9533, Std: 0.0340
8-bit Precision - Mean Accuracy: 0.9533, Std: 0.0340
4-bit Precision - Mean Accuracy: 0.9467, Std: 0.0542
2-bit Precision - Mean Accuracy: 0.8200, Std: 0.0957

Model: k-NN
Full Precision - Mean Accuracy: 0.9667, Std: 0.0298
8-bit Precision - Mean Accuracy: 0.9667, Std: 0.0298
4-bit Precision - Mean Accuracy: 0.9733, Std: 0.0249
2-bit Precision - Mean Accuracy: 0.7733, Std: 0.0611

Model: SVM
Full Precision - Mean Accuracy: 0.9667, Std: 0.0298
8-bit Precision - Mean Accuracy: 0.9733, Std: 0.0249
4-bit Precision - Mean Accuracy: 0.9733, Std: 0.0249
2-bit Precision - Mean Accuracy: 0.8200, Std: 0.0777
```

Analysis of Precision vs. Accuracy Trade-off

- **Higher Precision (8-bit)**: Maintains nearly the same accuracy as the original data.
- **Moderate Precision (4-bit)**: Slightly reduces accuracy, though acceptable for less sensitive applications.
- **Low Precision (2-bit)**: Leads to noticeable accuracy degradation, demonstrating a limit to how low the data precision can be reduced before model effectiveness is impacted.





Quantized Gradient Logistic Regression has a slightly higher mean accuracy than the full-precision model and demonstrates less variability in its performance. This means that the quantized version is slightly more robust or generalized, possibly due to regularization effects induced by quantization.

```
Full-precision Logistic Regression - Mean Accuracy: 0.9400, Std: 0.0442
Quantized Gradient Logistic Regression - Mean Accuracy: 0.9533, Std: 0.0163
```