Assignment 2

1. Training set (80%)

The choice is because of the learning capacity, model complexity and stability.

Learning Capacity: The training data should have large portion for model to learn from data. Thereby it allows the model to learn better representations and capture the underlying patterns in data.

Model Complexity:

Lacking training data paves the way to overfitting, Therefore, the larger data set prevents model from remembering the data and allows the unseen data to be generalized properly.

Stability:

Better convergence and potentially higher accuracy are important. Existence of more samples in the training set makes the mode’s training process more stable.

Validation Set (10%)

* Hyper parameter tuning: Validation set is used to fine tune hyper parameters such as learning rates, regularization strength and model architectures. It helps to examine with different configurations without contaminating the test set.
* Monitoring and Training : During training, you can regularly evaluate the model's performance on the validation set to detect overfitting or underfitting early. This helps in making decisions about when to stop training or adjust the learning process.

Test Set (10%):

* Generalization Assessment: The test set serves as an independent evaluation of the model's performance. It simulates how well the model will perform on new, unseen data. It provides a realistic estimate of the model's real-world performance.
* Bias and Variance Evaluation: The test set helps assess whether the model suffers from high bias (underfitting) or high variance (overfitting). By evaluating its performance on an unseen dataset, you can gauge how well it generalizes.