ACTIVIDAD #3

Tipo actividad: Reading comprehension "Train vs. Validation vs. Test set" and Kahoot activity

9) Socialize key words: "Train/Test Split and Cross Validation"

- Training Set: Set of data used to train and make the model learn hidden features/patterns.
- Validation Set: Separate set of data used to validate the model performance during training and tune hyperparameters.
- Test Set: Separate set of data used to test the model after completing the training and provides an unbiased final model performance metric.
- Overfitting: When the model becomes too good at classifying samples in the training set but fails to generalize to unseen data.
- Hyperparameters: Parameters that are set prior to training and affect the learning process of the model.
- Model Performance: Metric used to evaluate how well the model performs in terms of accuracy, precision, etc.
- Dataset Split: Process of dividing the dataset into training, validation, and test sets to judge the true model performance.

10) Reading comprehension activity: "Train/Test Split and Cross Validation"

Reading: "Train vs. Validation vs. Test set"

For training and testing purposes of our model, we should have our data broken down into three distinct dataset splits.

The Training Set
It is the set of data that is used to train and make the model learn the hidden features/patterns in the data.

In each epoch, the same training data is fed to the neural network architecture repeatedly, and the model continues to learn the features of the data.

The training set should have a diversified set of inputs so that the model is trained in all scenarios and can predict any unseen data sample that may appear in the future.

The Validation Set

The validation set is a set of data, separate from the training set, that is used to validate our model performance during training.

This validation process gives information that helps us tune the model's hyperparameters and configurations accordingly. It is like a critic telling us whether the training is moving in the right direction or not.

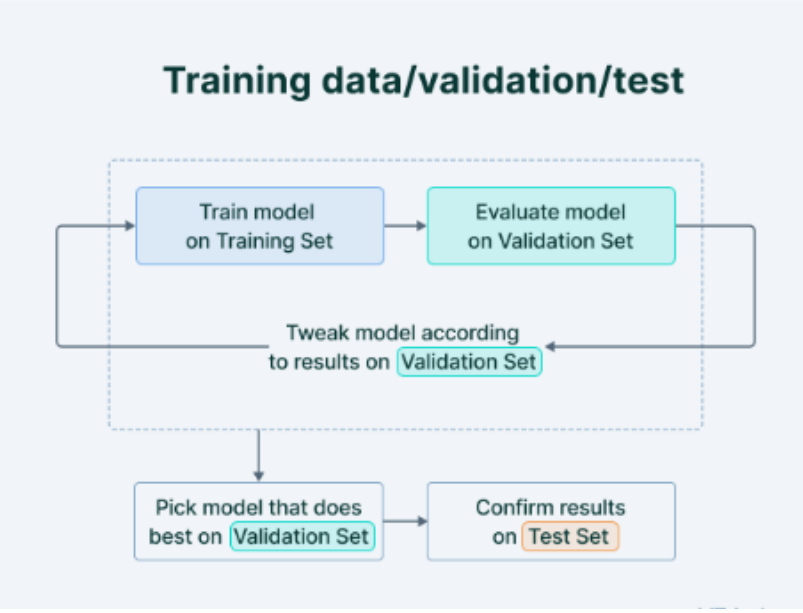
The model is trained on the training set, and, simultaneously, the model evaluation is performed on the validation set after every epoch.

The main idea of splitting the dataset into a validation set is to prevent our model from overfitting i.e., the model becomes really good at classifying the samples in the training set but cannot generalize and make accurate classifications on the data it has not seen before.

The Test Set

The test set is a separate set of data used to test the model after completing the training.

It provides an unbiased final model performance metric in terms of accuracy, precision, etc. To put it simply, it answers the question of "How well does the model perform?"



How to split your Machine Learning data?

The creation of different samples and splits in the dataset helps us judge the true model performance.

The dataset split ratio depends on the number of samples present in the dataset and the model.

Some common inferences that can be derived on dataset split include:

- If there are several hyperparameters to tune, the machine learning model requires a larger validation set to optimize the model performance. Similarly, if the model has fewer or no hyperparameters, it would be easy to validate the model using a small set of data.

- If a model use case is such that a false prediction can drastically hamper the model performance–like falsely predicting cancer–it's better to validate the model after each epoch to make the model learn varied scenarios.

- With the increase in the dimension/features of the data, the hyperparameters of the neural network functions also increase making the model more complex. In these scenarios, a large split of data should be kept in a training set with a validation set.

There is no optimal split percentage.

One has to come to a split percentage that suits the requirements and meets the model's needs.

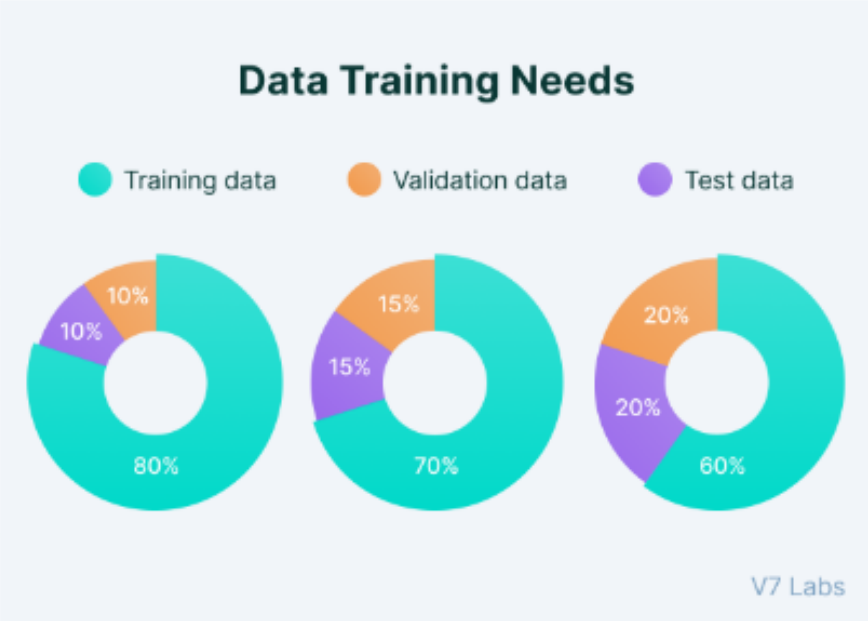
However, there are two major concerns while deciding on the optimum split:

If there is less training data, the machine learning model will show high variance in training.

With less testing data/validation data, your model evaluation/model performance statistic will have greater variance.

Essentially, you need to come up with an optimum split that suits the need of the dataset/model.

But here's the rough standard split that you might encounter.



Adapted from: https://www.v7labs.com/blog/train-validation-test-set#train-vs-validation-vs-test-set

11) Kahoot activity.

LINK to play online: https://create.kahoot.it/share/train-vs-validation-vs-test-set/d91ccd68-13bf-43ec-8a8a-e2af12d0fe1f