## L05 Reflective Journal

**Reflection on Learning**

* Today, we engaged in an exercise that saw us tackle image processing and classification by leveraging the CIFAR-10 dataset and the SVN algorithm.
* After installing and importing the requisite libraries that we would be implementing, we loaded our dataset and began cleaning/preprocessing. We went about this by converting our images to grayscale, followed by flattening the images and the labels immediately afterwards. We inspected our progress by generating a sample image which appeared to be in line with our expectations.

A close-up of a sample image

Description automatically generated

* The largest hurdle faced was with the unpredictability of the model training performance; as we did not realize until an incredibly late stage that the size of the dataset would prove to be problematic, at which point we reduced the size of the training and test sets each by an order of magnitude.

##### **Responses to Lab Questions**

* What is SVM? Why use SVM? What does ‘SVC(kernel=’lineal’)mean?
  + Support Vector Machine is a supervised machine learning algorithm that classifies data by finding an optimal line that maximizes the distance between each class in multi-dimensional space.1
  + SVM is widely used for classification purposes and is notable for its efficacy in high dimensional spaces.2
  + ‘SVC(kernel=’lineal’) means that we are specifying the use of a **lineal kernel**, as opposed to **polynomial or radial basis (RBF) kernel.** The use case for a lineal kernel is such where the data is separable by a straight line or hyperplane, as we have with CIFAR-10.

##### **Critical Analysis**

* Despite the numerous hangups with both provided Jupyter notebooks, I was able to get each working eventually.
* My largest takeaway from this exercise was in contrasting how I used my own dataset for one, vs using a ‘canned’ dataset like CIFAR.
  + When using a dataset as well-traveled as CIFAR, we inherit the benefits from the efforts of the community. The data has already been standardized, benchmarked, and vetted.
  + In contrast, using our own dataset means that these pre-processing duties fall entirely onto us.
  + I think that as part of the learning process, using CIFAR, Iris, or similar datasets is key to reliably understanding how model training works; whereas custom datasets can introduce complexity that may require attention paid to other areas like data collection & cleaning.

##### **References**

1. *What is Support Vector Machine? | IBM*. (n.d.). https://www.ibm.com/topics/support-vector-machine
2. *1.4. Support vector machines*. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/svm.html