The Recommendation Model

Using distance to calculate the relationship between data points is a common method in machine learning. Distance functions are a derivative of the Pythagorean Theorem. There are multiple distance metrics that can be used, however, in this project I chose to use Euclidean distance as a way of evaluating the relationship between data points. What Euclidean distance does is it calculates the distance between two points on a plane. The Euclidean distance between two vectors provides a measure of how similar or different the vectors are from one another. The closer two points are to one another, the more similar they are.

In regards to a machine learning application, this is useful because Euclidean distance is used in clustering machine learning applications, such as K nearest neighbors. The difference between using KNN and just using Euclidean distance is that in the case of this project, the model is not actively learning. It is simply doing calculations based on the distance of data points to generate recommendations. Had there been more time and had we not needed to scale down our project due to the loss of a third partner, implementing a “like” and “dislike” feature to further train the model and refine/improve song recommendations would’ve been the next logical step.

This model is evaluating different aspects of the music in relation to the artist that a user enters. The model is looking for songs that will have similar features to the song entered by the user by calculating the Euclidean distance between those song feature data points. Creating a working recommendation system was difficult because a lot of the data is in the form of strings rather than integers or floats. This doesn’t work well for algorithms that are trying to perform mathematical computations. For this reason, it was necessary to use One Hot Encoder for the genre column of the music. One Hot Encoder allows string values to be converted or “encoded” as numbers that a model can read and evaluate. It was also important to standardize and drop certain columns that contained string values before processing the distance.