Data preprocessing included selection and cleaning. Data was selected from the US Census Bureau website data portals. Specifically using the American Community surveys. Data was cleaned in excel, and in Pandas. It was cleared from 0 values, ended up eliminating three rows. In order to focus on our target question we chose an 85% threshold access to Internet based on the income brackets (under 10k to over 75k).

85% with internet access was chosen as a group, due to suitability within the group for both the ML and database management. We wanted all of our data to align with each other. Feature selection was determined by the data we had. We wanted to ensure consistency across the data streams, so to keep that (using counties in US and Puerto Rico), we were somewhat limited in features. We ended up with types of computer/income brackets and internet access. We ended up using desktop/laptop as a limiter, again to ensure compatibility throughout the group.

I chose income bracket with internet (85% or more) as the target variable. I could have used without at 15% as an option, but again, was going for alignment across the group.

Logistic regression (LR) was initially chosen, which was problematic until we chose our threshold. At that point, I have moved to unsupervised learning (UL). Using the UL showed trends from the under 10k to over 75k in the ratio of those with and without increasing as the income increased.

LR also showed the increase in internet access as the income increase, with the income bracket of 35k to 50k being a linchpin between the “haves and have nots.” Our accuracy scores for at 0 (access under 85%) was high at the under 10k and our accuracy score for 1 (access above 85%) at over 75k was also high. Accuracy score at the 35k to 70k showed 50% accuracy at the 1 and 0, indicating it was the turning point for the access availability.

The confusion matrix showed similar trends.