**CSCI – Machine Learning**

**William Diment**

This is my submission for the final project in CSCI – Unsupervised Machine Learning. In this submission, we take publicly accessible voter data for Boulder County and see if we can train a logistic regression model on the voter data. Then, using that data, we will attempt to predict the results of the 2024 Presidential Election in Boulder County. We will be able to directly compare our results from the training data to the test data, and obtain detailed information regarding the accuracy/error from our models. We begin by obtaining the 2020 voter registry from Boulder County, located here:

<https://bouldercounty>.gov/elections/maps-and-data/data-access/

The voter registry in total, split across two parts, contains the registration details of 249,361 registered voters, where each registered voter is a row – the index for the row is the voter ID. There are 82 columns, and the columns we are concerned with are: Voter\_Status, Party, Gender, Birth\_Year, Precinct\_Code – note, there is an additional column of precinct\_name that is a duplicate of the precinct\_code column.

We will need to drop a lot of extraneous columns from this data set in order to actually get the information we need, this is the bulk of the EDA that needs to be done. This is done in a function called Clean Voter Registration – in it, we select only the columns we want, replace string values with signifiers, and then drop any party that is not classified as being Democratic or Republican. We then concatenate the two separate voter lists in order to get the full voter list

The actual Machine Learning model that we have chosen is Logistic Regression. This is actually somewhat against what I was thinking I would pick in this given situation – I wanted to use a classifier.