**Dataset 1: The Iris Dataset:**

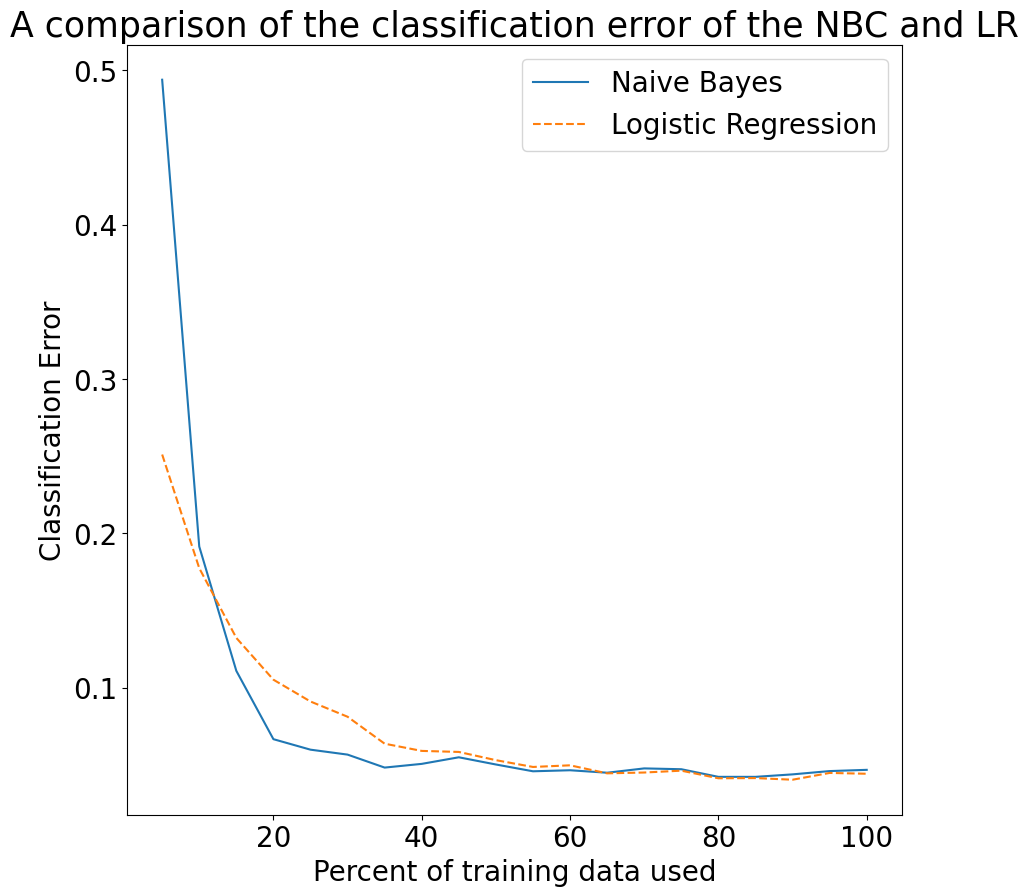
The Iris Dataset contains data regarding the petal and sepal length of three different species of the Iris flower (Setosa, Versicolour, and Virginica), stored in a 150x4 numpy.ndarray. By obtaining the first five rows of the dataset, we can see that there are no NaN values. This is confirmed by obtaining a concise summary of the dataframe by using the info function. Following the initial data analysis, we standardised the data which has the goal of rescaling the data to achieve a mean value of 0 and a standard deviation value of 1 for each feature. We had shuffled the data and split it into train and test categories with the ratio 80:20 respectively and created the labels (x\_train, x\_test, y\_train and y\_test).

**Dataset 2: Voting Dataset:**

The Voting Dataset contains data regarding the votes from democrats and republicans on 16 different topics. Each column is supposed to have 435 entries, but some columns appear to contain less entries which indicates the presence of NaN values. To solve this, we dropped the observations with NaN values. Furthermore, we sampled 100 random observations to comply with the exercise’s guide. We also converted text categorical features into real numbers by using the Ordinal Encoder class. We abstain from scaling this dataset because of its binary nature. The data is shuffled and split it into train and test categories with the ratio 80:20 respectively and the labels (x\_train, x\_test, y\_train and y\_test) are created.

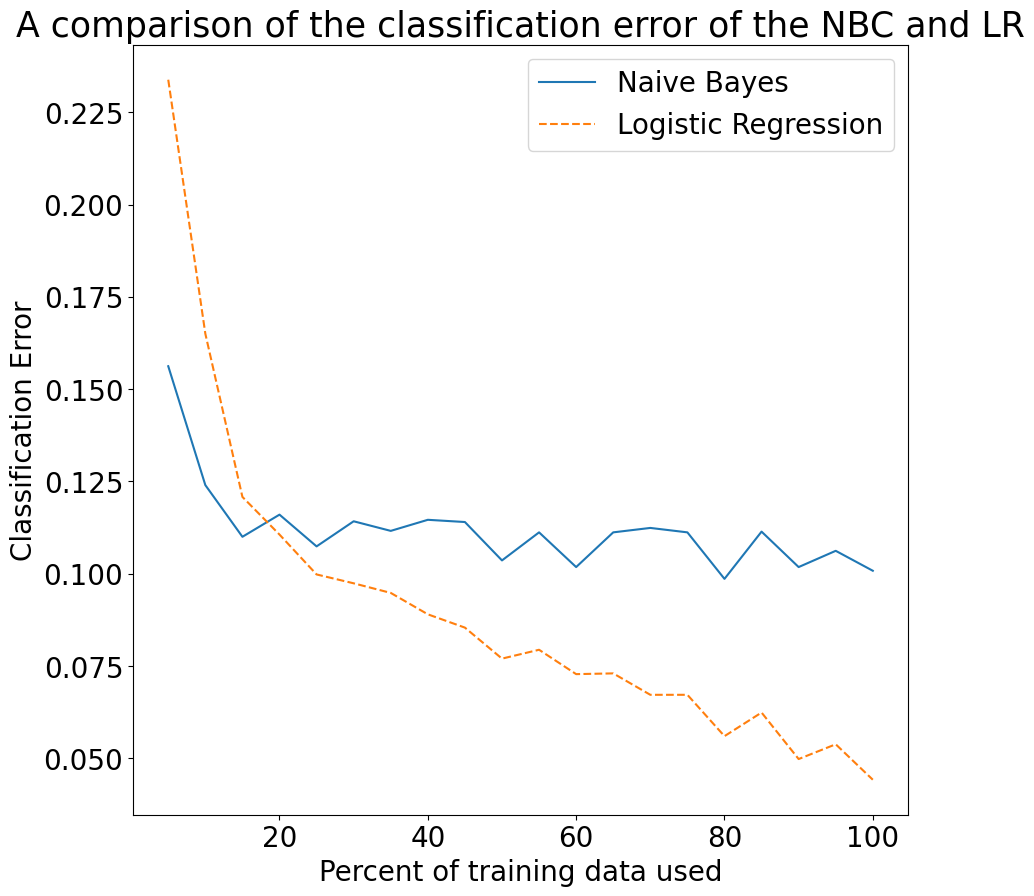
**Observations of the comparisons:**

**Iris dataset:**



The relationship between NBC and Logistic Regression meets our predictions: the naïve bayes classifier initial has a higher asymptotic error which quickly falls and generally coincides with the performance of the logistic regression algorithm. At smaller datasets, the logistic regression algorithm’s performance sometimes does not meet the performance of the naïve bayes classifier which could be because the logistic regression performs much better in larger datasets.

**Voting dataset:**



The logistic regression algorithm has a higher asymptotic error but the naïve bayes classifier quickly converges to a higher asymptotic error. Therefore, as the proportion of training data used increases, the naïve bayes classifier falls in performance and is overtaken by the discrimination logistic regression.

Both graphs also illustrate that the naïve bayes generative model reaches its asymptotic error at a much faster rate than the discriminative model when using logarithmic data, as we have done.