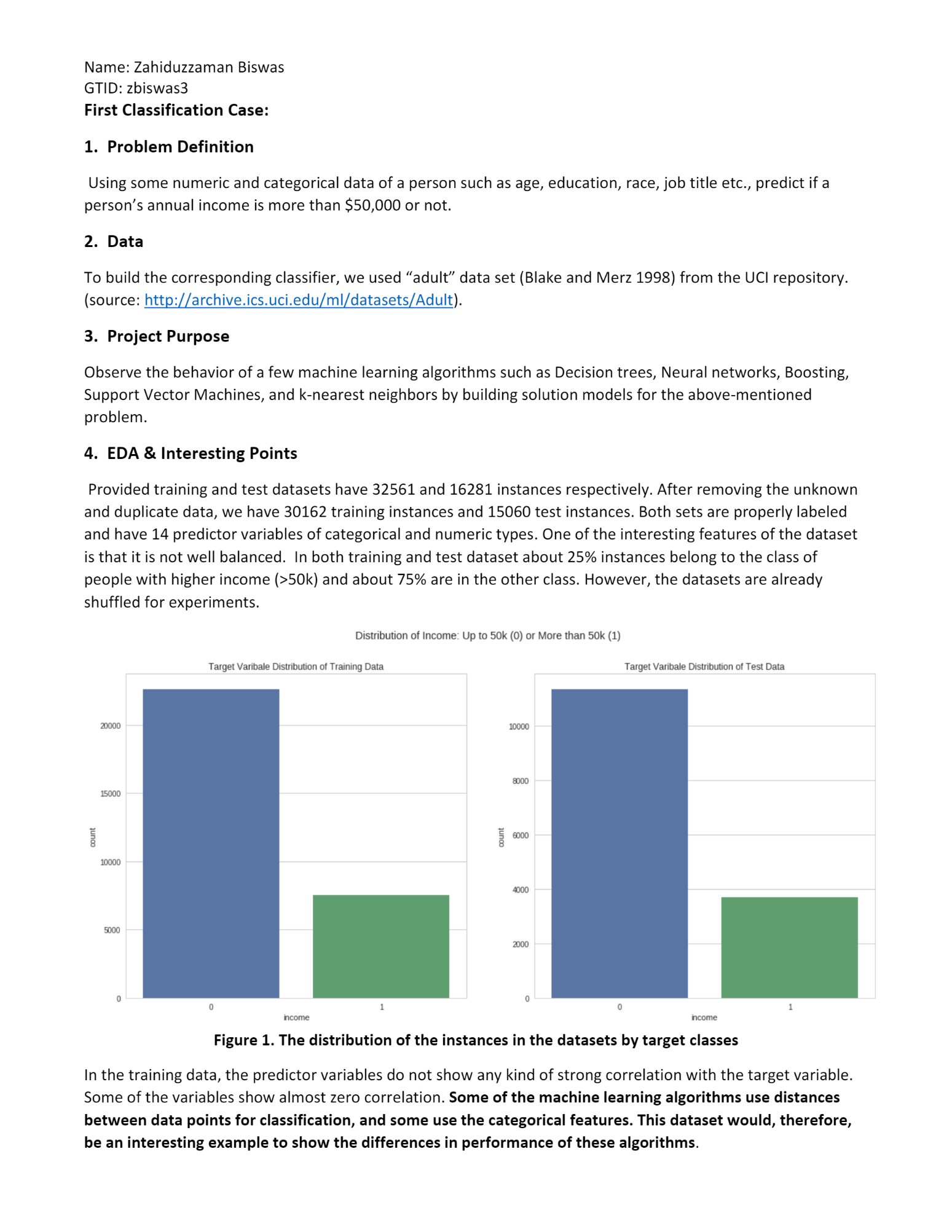
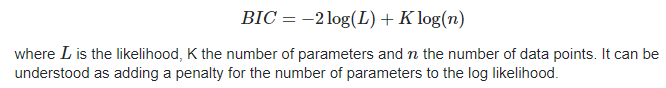
**Adult Income Dataset**: In the last decade, the global population living in poverty (defined as living with less than $2 per day, 1985 prices) has decreased dramatically from 80% in 1920, to 50% in 1970 to 10% in 2015 [1]. In order to determine the policy changes, US government primarily uses census data. Adult income dataset is a subset from the 1994 US Census. It provides the data on education, heritage, work status and age (among others) to classify whether income is above $50,000 per year or not. US governments can use this data to determine the most impactful factors on individual income in the USA. The dataset consists of 14 socio-economic attributes of categorical and numeric type; and 32,561 instances which is sufficiently large enough to find statistical significances of different factors on the income. One of the interesting features of the dataset is its uneven distribution. In terms of machine learning, it poses a challenge for classifiers to learn from the dataset.



E-M Clustering:



Under the hood, a Gaussian mixture model is very similar to *k*-means: it uses an expectation–maximization approach which qualitatively does the following:

1. Choose starting guesses for the location and shape
2. Repeat until converged:
   1. *E-step*: for each point, find weights encoding the probability of membership in each cluster
   2. *M-step*: for each cluster, update its location, normalization, and shape based on *all* data points, making use of the weights

The result of this is that each cluster is associated not with a hard-edged sphere, but with a smooth Gaussian model. Just as in the *k*-means expectation–maximization approach, this algorithm can sometimes miss the globally optimal solution, and thus in practice multiple random initializations are used.

