1. **Introduction**

Solar photovoltaics is one of the most promising approaches for renewable energy, which is a method of converting the energy in the sun’s rays into electrical energy. As the development process for most modern solar cells is expensive and rigid, a machine learning process is desirable to rapidly identify molecular features that may determine their potential efficiency, measured by the difference in energy between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO). This report considers and evaluates the feasibility of several most popular machine-learning algorithms in this specific scenario.

1. **Data Description**

We use a subset of data from The Harvard Clean Energy Project, which features 1,000,000 molecules for training and 800,000 for testing, given in SMILES format. The response variable we attempt to predict is the difference in energy between the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO). The empirical distribution of energy gap is displayed in Figure 1. As shown, the distribution is unimodel and roughly symmetric, it is thus reasonable to conclude that it follows normal distribution. Among the 1,000,000 observations, three negative records were observed. Since energy gap is defined as the difference between highest and lowest occupied molecular orbital, the negative values were considered illogical values and were hence removed from subsequent analysis. 256 binary features were extracted with RDKit to predict target energies, comprising a 1,000,000 x 256 sparse matrix.

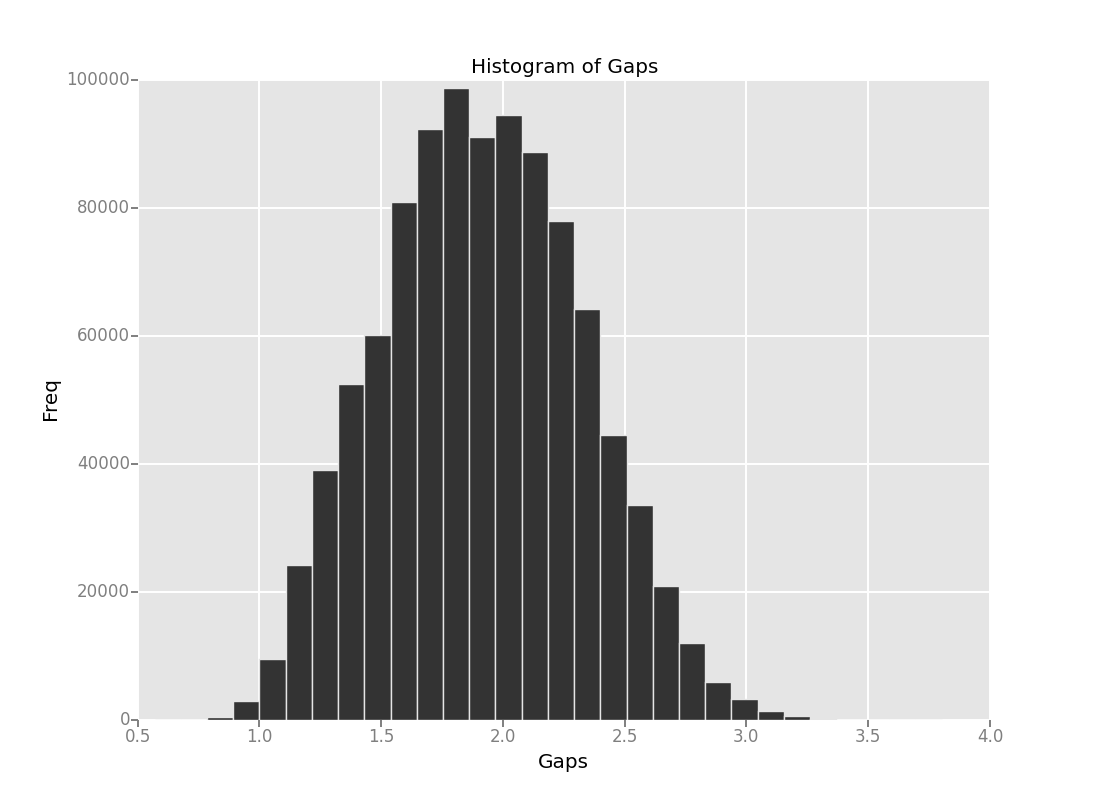


Figure 1: Histogram of Outcome Variable: Gaps