1. Problem Introduction

The field of Nanotechnology has been experiencing substantial growth rates for a few years now. While nanotechnology is currently being applied to a broad range of problems across different fields such as Energy, Electronics, Medicine, etc. I wanted to get a glimpse of what nanotechnology problems are currently being studied and came across the problem of toxicity. Since nanoparticles are extremely small, especially those with sizes less than 100 nanometers. These nanoparticles can easily enter and lodge themselves in organs within the human body. Consequently, causing damage or even killing cells within the body also known as cytotoxicity.

For this project I will be aiming to develop a few different classification Machine Learning models, assessing their performance through different metrics and optimizing to achieve the best possible results.

2. Data Introduction

For this project I will be working with the Nanoparticle Toxicity Dataset found on Kaggle. This dataset contains about 800 records which are composed of several features including: core size, hydrodynamic size, surface area, surface charge, electric charge, dosage, exposure time, number of oxygen atoms, and energy (possibly surface energy). The dataset also contains a class which takes values of Toxic and nonToxic. I have a few concerns about this dataset, one of them being that it does not contain any information regarding units of measurement. Another concern is that this dataset can raise possible misinterpretation, mainly due to the classification of Toxic or nonToxic. This classification can mean different things given a different context, for example a toxic nanoparticle may only be toxic to a specific cell type such as cancer cells in which case this classification would be desired. Additionally, if the nanoparticle is toxic to benign cells within the body, then this is an unwanted result.

3. Data-Preprocessing

This dataset is relatively clean, as expected from a Kaggle dataset. I began by assessing the missing value count, which was 0 in this case. The only issue I ran into with this dataset was that many records were duplicated; after removing the duplicates only 1/3 of the original data was left. I suspected that this would create significant issues in the distribution of the classes as confirmed later by the visualizations.

This dataset contains one categorical feature which signifies the Chemical compound of the nanoparticle, I opted to use this feature for my models rather than simply removing it. Thus, I had to perform one-hot encoding in order to make them usable predictors in the models.

4. Visualization & Modeling

5. Evaluation