**David Montes de Oca Zapiain**

ME PhD 2019 Georgia Institute of Technology

**Citrine Data Challenge Submission**

The objective of this work is to leverage data-science and data-analytics to predict the stable binary compounds an element pair will produce when mixed; a renowned thermodynamics problem.

In order to limit the effect of other phenomena (i.e. heat transfer, time, etc.) and to focus on the effect the different features provided for each element pair have on generating a stable binary compound the following assumptions were used:

* The system is a closed system with negligible heat transfer (adiabatic)
* The kinetic and potential energy of the system are neglected since it is assumed that the results provided are in steady state.
* The results of each component of the stability vector are assumed to be independent of each other.

The data provided has 2572 pairs of elements with 98 features per pair of elements. The objective is to leverage this data to produce a robust model capable of predict the stable binary compounds that each element pair will produce. The stable compounds are a discretization of the 1D binary phase diagram at 10% intervals, where a 1 indicates a stable compound at that concentration of elements and a 0 indicates that no stable compound has been produced. It is also important to notice that the first and last column of the stability vector are always stable (all have a 1) since these columns represent a 100% compound of each of the elements in the pair, which are inherently thermodynamically stable. Therefore, there is no need to build a model for these components of the stability vector.

-trying to build a model that predicts all the stability vector at the same time using a multi-class classifier is not good because the number of classes you have is extremely big and the data provided is limited, therefore will be an under fit.

-Therefore, it might be much smarter to systematically break down the problem.

Having said that, a first logical step is to investigate the columns of the stability vector that characterize the binary compounds produced by each element pair and determine the number of stable compounds each pair produces. An efficient way to do this is by plotting a histogram of the number of stable compounds each element pair produces as shown in Figure 1.



**Figure 1. Number of Stable Compounds per element pair.**

From Figure 1 it can be clearly seen that the number of element pairs that produce no stable compounds at all is roughly half of all the element pairs. A model that determines whether the input element pair will procure at least one stable compound is a good first step to address this this complex problem since it reduces a multi-class problem to a binary classification problem. In other words, this first model is nothing more than a binary classifier that will leverage the features of the element pairs to discern which ones of them yield a stable compound.

A model built by taking into consideration the features that have highest influence/impact on the output yields more robust and accurate results. Therefore, the first step in building this model is pre-processing the features provided to identify the most relevant features.

I will evaluate the performance of this first model with the help of a confusion matrix and the area under the ROC curve, which is an extremely suitable metric for assessing the performance of binary classifiers.