Model to Predict the Critical Temperature of Superconductors

This chapter builds a regression machine learning model to predict the critical temperature of superconductors. The features for this dataset were derived based on the following superconductor properties:

- Atomic mass

- First ionization energy

- Atomic radius

- Density

- Electron affinity

- Fusion heat

- Thermal conductivity

- Valence

And for each property, the mean, weighted mean, geometric mean, weighted geometric mean, entropy, weighted entropy, range, weighted range, standard deviation, and weighted standard deviation are extracted. Thus, this results in a total number of 8 x 10 = 80 features. In addition to this, a feature that contains the number of elements in the superconductor is added to the design matrix. The predictor variable is the critical temperature of the superconductor. Hence, the dataset has a total of 81 features and 21,263 rows.

This dataset is made available by Kam Hamidieh of the University of Pennsylvania and submitted to the UCI Machine Learning Repository. The goal of this section is to demonstrate delivering an end-to-end machine learning modeling pipeline on Google Cloud Platform.

The Modeling Architecture on GCP

The goal of this end-to-end project is to demonstrate building a large-scale learning model on GCP using the components already discussed in this book. The modeling architecture is illustrated in Figure 44-1. Let's briefly explain the connections:

1. Stage the raw data on GCS.

2. Load data into BigQuery for analytics.

3. Exploratory data analysis.

4. Large-scale data processing with Dataflow.

5. Place transformed training and evaluation data on GCS.

6. Train the model on Cloud MLE.

7. Place the trained model output on GCS.

8. Deploy the model for inference on Cloud MLE.