**Project Proposal**

ch3470 Chenyu Huang

* **Background**
  + [ClinVar](https://www.ncbi.nlm.nih.gov/clinvar/) is a public resource containing annotations about human genetic variants. These variants are (usually manually) classified by clinical laboratories on a categorical spectrum ranging from benign, likely benign, uncertain significance, likely pathogenic, and pathogenic. Variants that have conflicting classifications (from laboratory to laboratory) can cause confusion when clinicians or researchers try to interpret whether the variant has an impact on the disease of a given patient.
* **Dataset**
  + Clinvar
    - Refer to <https://www.kaggle.com/kevinarvai/clinvar-conflicting>
    - Manual Classification Label: Benign, Likely Benign, VUS, Likely Pathogenic, Pathogenic.
    - Merged Classification Label[1]: 1 -> conflicting, 0 -> not conflicting.
    - 65188 variants contained
    - 40 features (numerical and categorical) after cleaning
  + Data Cleaning
    - Calculate the rate of meaningless contents (e.g. ‘nan’s, empty or ‘not\_specified’) within each column and sort the rate in descending order.
    - Double-check the elements in the columns with python ‘set’ before eliminating them.
    - Eliminate those columns with a rate higher than 99%
* **Questions**
  1. Estimate how many variants will have conflicting classifications, why are they considered to have conflicting classifications?
     + Methodology: Logistic Regression
  2. Evaluate the importance of all these features, find the most important one and analyze why is it so important.
     + Methodology: Decision Tree, Bootstrap (and probably MLE)
  3. Redo Question 1 with classification label unknown, compare the 2 results.
     + Methodology: K-means Clustering

**Appendix – Data Cleaning Process**



