**Human Genetic Variants Classification**

ORCA 4500 Foundation of Data Science, Project Report

Chenyu Huang

Team #5

* **Contents**
  + About the Dataset
  + Problem Statement
  + Classification
    - Logistic Regression
    - Decision Tree
    - Random Forest
    - Adaboost
  + Feature Engineering
    - Random Forest
  + Clustering
    - K-means
* **About the Dataset (Clinvar)**
* **Questions**
  1. Estimate how many variants will have conflicting classifications, why are they considered to have conflicting classifications?
     + Methodology:
       - Logistic Regression
       - Decision Tree
       - Random Forest
       - Adaboost
  2. Evaluate the importance of all these features, find the most important one and analyze why is it so important.
     + Methodology:
       - Random Forest
  3. Redo Question 1 with classification label unknown, compare the 2 results.
     + Methodology
       - K-means Clustering