I am analyzing genetic variant classification data provided by three clinical labs and collected and curated by ClinVar. From their website, "ClinVar is a freely accessible, public archive of reports of the relationships among human variations and phenotypes, with supporting evidence. ClinVar thus facilitates access to and communication about the relationships asserted between human variation and observed health status, and the history of that interpretation. ClinVar processes submissions reporting variants found in patient samples, assertions made regarding their clinical significance, information about the submitter, and other supporting data. The alleles described in submissions are mapped to reference sequences, and reported according to the HGVS standard. ClinVar then presents the data for interactive users as well as those wishing to use ClinVar in daily workflows and other local applications. ClinVar works in collaboration with interested organizations to meet the needs of the medical genetics community as efficiently and effectively as possible."

The dataset I am using is found on Kaggle as "Genetic Variant Classifications". This data set contains an estimate 65200 records with 46 variables measured per record (65200x46 dataframe). Each record represents an allele- a variant form of a given gene. 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.

I am only provided the information that either all participating labs classified the variant in the same category (CLASS column is 0) or at least two of the labs disagree on the category of that variant (CLASS column is 1). Thus the data analysis of this dataset will focus on a binary classification, while examining relative variables that may be linked to each classification case.

The binary variable of interest, CLASS will be analyzed across several columns (variables) to determine patterns and relationships between classifications and other measured quantities. Both binary cases- classification agreements and disagreements- will be analyzed.

The application of this analysis aims to benefit ClinVar directly, the participating labs, and the broader medical genetics community. My analysis seeks to determine which feature variables have the biggest impact on the target variable.

If my analysis can find a relationship between one or more of the 45 variables and the binary CLASS classification, the results could help future labs efforts in examining genetic variations in patients and having more consistent classifications as other labs.