Dataset appears to be a single nucleotide polymorphism (SNP) datasetSNPs (Single Nucleotide Polymorphisms) are bi-allelic by definition. Bi-allelic variants(DNA sequences, ) , such as SNPs, can only have two possible alleles (nucleotide bases) at a specific genomic position. However, when considering the genotype of an individual at that position, there are three possible scenarios:

1. Homozygous for the major allele (0)
2. Heterozygous (1)
3. Homozygous for the minor allele (2)

For example, let's consider a SNP at a particular position on chromosome 1, where the two possible alleles are A (major allele) and G (minor allele).

1. Homozygous for the major allele (0): An individual with the genotype AA at this SNP position is homozygous for the major allele (A). This genotype is typically represented as 0.
2. Heterozygous (1): An individual with the genotype AG or GA at this SNP position is heterozygous, carrying one copy of the major allele (A) and one copy of the minor allele (G). This genotype is typically represented as 1.
3. Homozygous for the minor allele (2): An individual with the genotype GG at this SNP position is homozygous for the minor allele (G). This genotype is typically represented as 2.

For example, The string "rs6656401" is a unique identifier assigned to a specific single nucleotide polymorphism (SNP) in the dbSNP database maintained by NCBI. This identifier allows researchers to easily access information about that particular SNP, including its chromosomal location, the two possible nucleotide alleles observed at that position, the allele frequencies in different populations, any known functional annotations or effects, and relevant literature references. The "rs" prefix stands for "reference SNP cluster ID." Using standardized SNP identifiers like rs6656401 facilitates data integration and collaboration in genetic and genomic studies by providing an unambiguous way to refer to a specific SNP across different databases and publications.

The dataset contains genotype information for 300,000 SNPs across 500 individuals, with each SNP having three possible genotypes (0, 1, or 2, representing the number of minor alleles). One-hot encoding was applied to represent these three genotypes as separate binary features, resulting in 900,000 features (300,000 SNPs × 3 genotypes per SNP).