**Capstone Project – Data Storytelling**

The 2D features are also known as Surface Property Maps and are just numbers representing the difference between the surface map of 'normal' amino acid (wild-type) and the mutated one. The 3D features, also known as Structure Distance Maps, are representing difference in magnitude of the distance changes in the 3D structure. This kind of multidimensional data is very hard to visualize for any specific feature correlation. We can try feature selection and dimensionality reduction on the data set and then try to visualize the key components. These features are bound to have highly correlated variables and dropping the highly correlated variables is part of feature selection.

Since there are multiple sites (1 pt to 6 pt) of mutations, we can see the distribution of the data set among these different types by plotting a histogram. The histogram shows that there are close to 31000 entries for the 2 pt mutation. The 3 pt and 1 pt have approx. 110 and 60 entries. The rest are represented by 30, 8 and 3 entries for 4 pt, 5 pt, and 6 pt mutations. Though, the data set is skewed with 31008 'inactive' and 151 'active' entries and we have both classes represented in all types of mutations. This distribution also tells us which ‘distance feature’ we can expect to be more prevalent.

Since, there are 31000 2 pt mutations; the ‘distance1’ will have the most number of entries. The distance features were plotted to observe the range of the differences between adjacent mutations. The 'distance1' feature which contains the difference in position between the first and second mutations shows that half of them lie between 0-90.

The 'distance2' feature represents the distance between the second and third mutations. This shows an even narrower range - most values are between 0-10 i.e., the adjacent mutations are really close most of the time. The 'distance3', 'distance4' and 'distance5' features were also seen to give a similar result. It is safe to say that in multiple mutations (more than 2 pt) adjacent mutations are mostly located close together - within a distance of 0-20.

The distributions of the active and inactive classes among the different mutations are in the table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mutations Count | Inactive | Active | Total |  |
| 1 | 55 | 8 | 63 |
| 2 | 30877 | 65 | 30942 |
| 3 | 49 | 63 | 112 |
| 4 | 24 | 7 | 31 |
| 5 | 2 | 6 | 8 |
| 6 | 1 | 2 | 3 |