Data Analysis

**GRD - Genomic based 16s Ribosomal RNA Database has been used and the analysis is as follows-**

* The database consists of 13252 sequences.
* The average length of each sequence is 1500 base pairs. However, it ranges from 65 bp to 3606 bp(outliers).
* The combined length of all the sequences is 20144850 base pairs.
* The nucleotides present are as follows-

**{'C': 4572015 'A': 5065337 'T': 4178530,**

**'G': 6312765 'R': 43 'S': 10,**

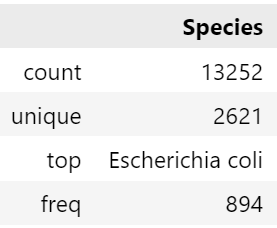
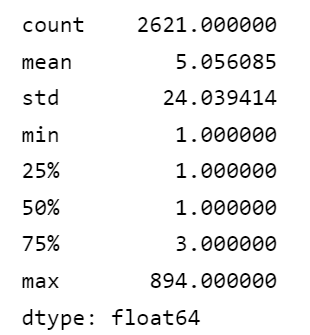
**'K': 16 'N': 16074 'M': 15,**

**'Y': 31 'W': 13 'V': 1}**

* The taxonomic classification for each sequence in the database is as follows:

Domain > Phylum > Class > Order > Family > Genus > Species > Strain

* Out of 13,252 sequences, 97 sequences contain IUPAC ambiguity codes (nucleotide symbols other than A,T,G,C)
* Out of these 97 sequences 71 sequences contain less than 50 codes and 19 sequences having codes between 630 to 910 codes. The remaining 7 sequences contain 83, 95, 115, 127, 203, 252, 384 codes resp.
* These 13,252 sequences, correspond to 2621 unique species.
* The 2621 unique species occur at an average frequency of 5 times in the database. However, the more than 50 percent, i.e. 1389 species appear only once.



* In its basic form, the neural network model gives the most accuracy which is around 83%.
* Controlling the probability threshold and increasing it to 80 percent, increased the accuracy to 93.7%. However increasing the probability threshold will decrease the total number of predictions made(recall) but will increase the reliability of the predictions(precision).