After the removal of homologous sequences in both positive and negative dataset using

CD-Hit, in order to carry out a binary classification between ubiquitination and non

ubiquitination sites, please accomplish the tasks described as follows.

1. **[Performance Comparison of Different Feature Encoding Methods] (50 points)**

You have learning a couple of feature encoding methods, such as One-hot encoding,

Amino acid composition (AAC), Amino acid pair composition (AAPC), Positional Weighted

Matrix (PWM), Position-specific scoring matrix (PSSM) and BBLOSUM62, from this course.

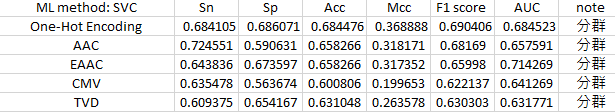
According to the k-fold cross-validation **(k should be larger than five)**, please do the

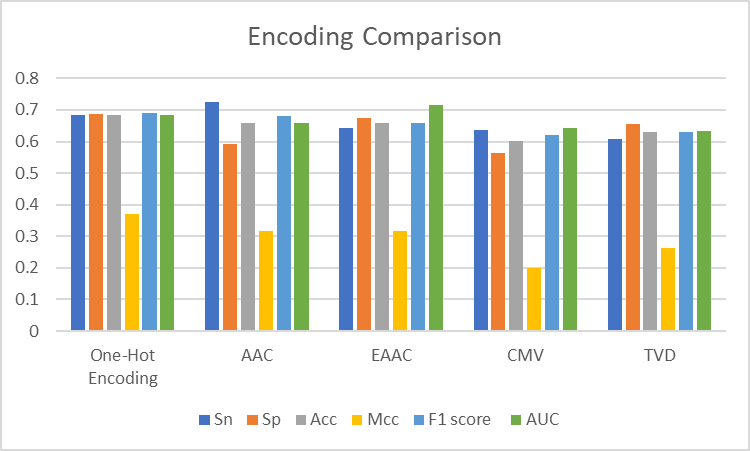
performance comparison between the models trained using different feature encoding

methods **(at least five feature types)**, based on a supervised learning method (e.g.

support vector machine). The comparison results should be provided in terms of a table

and bar-charts, as the example shown below.





2. [Performance Comparison of Different Supervised Learning Methods] (50 points) After the performance evaluation of different feature encoding methods, you can then try to consider different supervised learning methods (at least five classifiers) into the construction of predictive models using the best feature or hybrid features, with an attempt to identify the best model. The comparison results should be provided in terms of a table, bar-chart representation, and ROC curves, as the example shown below.

