# 2. Methods

## 2.1 Dataset

In this study, we utilized the Protein Folding Kinetics Database, PFDB dataset (<http://lee.kias.re.kr/%7ebala/PFDB>, accessed on 5 May 2020), which comprises 141 single-domain proteins without covalently bound prosthetic groups or disulfide linkages [1]. Among these proteins, 89 were classified as two-state (2S) and 52 as non-two-state (N2S) globular proteins. PFDB, being the most recently updated database with accurate ln(kf) values and folding types, serves as the gold standard for developing prediction models. Additionally, we collected protein sequences from the Protein Data Bank (PDB) website (<https://www.rcsb.org/>, accessed on 5 May 2020), using the PDB\_IDs listed in the PFDB dataset. Initially, we downloaded files in .fasta format, each corresponding to a PDB\_ID containing the protein sequence. Subsequently, we developed a Python program to read these files and extract the protein sequences based on the specified chains and residues in the PFDB dataset. Finally, we integrated these sequences into the existing PFDB dataset, leaving the other columns unchanged.

## 2.2 Feature Extraction

// Todo

## 2.3. Machine learning algorithms

// todo

## 2.4. Evaluation metrics

We employed a binary classification approach to predict the folding type, categorizing the elements into two distinct classes: 2S and N2S. Evaluating the outcomes of binary classification models is a crucial aspect of machine learning and computational statistics. Whenever researchers apply an algorithm to differentiate between two conditions (e.g., positive and negative) within a dataset, they can construct a contingency table known as a two-class confusion matrix. This matrix provides a comprehensive representation of the correctly predicted elements and those that were misclassified [2]. Among the positive instances, the elements that the algorithm accurately identified as positive are referred to as true positives (TP). Conversely, the positive instances that were incorrectly classified as negative are designated as false negatives (FN). Similarly, for the negative instances, those that are correctly labeled as negative are termed true negatives (TN), while the negative elements that are wrongly predicted as positive are called false positives (FP). To assess the performance of our classification model, you utilized five widely used evaluation metrics: Accuracy (ACC), Matthew's correlation coefficient (MCC), Precision (PRC), Sensitivity (SNS), and the F1-score (F1) [3]. The definitions of each metric are as follows:

AAC =

MCC =

PRC =

SNS =

F1 =

# 3. Results and discussion

// todo

References

1. B. Manavalan, K. Kuwajima, J. Lee, PFDB: a standardized protein folding database with temperature correction, Sci Rep 9 (2019) 1588
2. Wei Q, Dunbrack Jr. RL. The role of balanced training and testing data sets for binary classifiers in bioinformatics. PLoS ONE. 2013; 8(7):e67863
3. Luque A, Carrasco A, Martín A, de las Heras A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recogn. 2019; 91:216–31