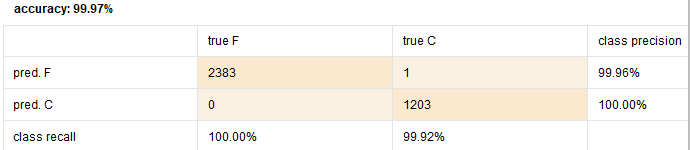
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Reported information | Test type | Decision Tree | Logistic Regression | Naïve Bayes | K-NN | Random Forest |
| Accuracy with default parameters | Entire dataset | 67.05 | 71.06 | 51.13 | 99.97 | 66.43% |
| 50% | 66.20% | 71.57 | 52.29 | 100 | 66.67 |
| Cross-validation | 66.55% +/- 0.38% | 70.53% +/- 1.55% | 50.96% +/- 2.01% | 62.67% +/- 2.34% | 66.24% +/- 0.50% |
| Accuracy with best parameters | Entire dataset | 67.58 | 72.04 | 51.13 | 99.97 | 66.77 |
| 50% | 66.7 | 71.97 | 52.29 | 100 | 67.28 |
| Cross-validation | 66.66% +/- 0.29% | 70.81% +/- 2.80% | 50.99% +/- 3.61% | 66.43% +/- 2.53% | 66.43% +/- 2.45% |
| List names of parameters | | Maximal Depth  Apply Pruning  Confidence  Apply Pre-pruning  Minimal Gain  Minimal leaf size  Minimal Size for split  Number of Pre-pruning | Reproducible  Use Regularization  Standardize  Non-neg Coefficients  Add intercept  Compute p-values  Remove collinear columns  Missing values handling  Max iterations  Max runtime Second | Laplace Correction | K  Weighted Vote  Measure Type  Numerical measure | # of Trees  Criterion  Maximal Depth  Apply Pruning  Apply Pre-pruning  Guess subset ratio  Voting Strategy |
| List selected best values of parameters (in the same order as in the list of names) | Entire dataset | 25  On  0.25  Off | On  Max Number: 2  Off  Off  Off  Off  Off  Off  Skip  10 | On | 1  On  Numerical Measure  Euclidean Distance | 4  Gain\_Ratio  10  Off  On  Minimal Gain: .01  Minimal Leaf Size: 1  On  Confidence Vote |
| 50% | 25  On  0.3  On  0.1  3  3  2 | On  Max Number: 6  Off  On  On  On  On  On  Mean Imputation  0 | On | 1  On  Numerical Measure  Euclidean Distance | 4  Gain\_Ratio  10  Off  On  Minimal Gain: .01  Minimal Leaf Size: 1  On  Confidence Vote |
| Cross-validation | 20  On  0.2  On  0.142  2  4  2 | Split on batch: Off  Leave one out: Off  Number of folds: 15  Sampling Type: Stratified Sampling  Use Local Random seed: Off  Enable parallel execution: On | Number of Folds: 15  Sampling type: Stratified Sampling | Number of Folds:11  Sampling Type: Linear Sampling  1  On  Numerical Measure  MaxProductSimilarity | Number of Folds:11  Sampling Type: Shuffled Sampling  10  Gain\_Ratio  10  On  Confidence:0.25  On  Minimal Gain: .01  Minimal Leaf Size: 1  On  Confidence Vote |

1. The models that I decided to use were K-NN and Random Forest. From what I found K-NN was pretty accurate to begin with so any small change in parameters would decrease the accuracy massively. I found if I changed the K value to any integer above 2 it would decrease the accuracy from almost perfect to about 67%. Numerical Measure also played a relatively big part in improving/decreasing accuracy. Changing the default value of Euclidian distance to Max Product Similarity lowered the value to about 66%. For Random Forest the values were much lower to start off but with some tweaking I could improve the values by a slight margin. Decreasing the number of trees tended to increase the accuracy. Turning off pruning also produced some favorable results.
2. Check Table Above
3. Between the three tests I believe that doing a split of the data into 50% and performing the model is much more beneficial than the other two tests. With my tests, most of the accuracies were the best in the 50% split.
4. For some of the algorithms it was definitely worth the time to play around with the values to improve the accuracy, but for most of the algorithms that were used for this project it was not worth the time. This was the case because for an algorithm like K-NN the accuracy is already high with default values, so trying to make the slightest improvement is a waste of resources. Another example is the Naïve Bayes, there is only one parameter that can be altered, and even changing that provided no significant improvement.
5. I believe that the accuracy of my K-NN model with Split data is sufficient for practical purposes. The K-NN has near perfect accuracy of 99.97% which makes this algorithm is very practical.
6. 
   1. This model would be suitable to identify proteins that crystallize and not crystallize. It is beneficial because it boasts an almost 100% accuracy to find crystallized proteins and boasts 100% accuracy for non crystallized proteins.