Assignment 1 The work is done by Group 3. The group members are: Zhou YANG, Chengran YANG, HARDYANTO PUTRA Antoni, Zhi CHEN and Zilin LUO.

Part I

We obtain the following results on the test data:

- Accuracy: 71.00%
- Precision: 66.27%
- Recall: 98.21%
- $F_1 = \frac{2*PR}{P+R} = 79.13\%$

Part II

According to this Wikipedia page [1], the 10-fold cross-validation randomly partitioned the original samples into 10 equal sized sub-samples. Of the 10 sub-samples, a single sub-sample is retained as the validation data for testing the model, and the remaining 9 sub-samples are used as training data. The cross-validation process is then repeated 10 times, with each of the 10 sub-samples used exactly once as the validation data. The 10 results can then be averaged to produce a single estimation. However, the standard 10-fold cross-validation can make the label distribution of the validation data very different from the training data, which motivates the *stratified cross-validation*. The stratified cross-validation requires the splitting of data to be governed by criteria such as ensuring that each fold has the same proportion of observations with a given categorical value, such as the class outcome value.

In this part, we use the following four classification methods: (1) SVM, (2) Random Forest, (3) Logistic Regression and (4) Deep Neural Network. Their results on both standard and stratified cross-validation are shown as follows.

SVM

Table 1: SVM results on the standard 10-fold cross-validation

k	**Acc**	**Prec**	**Reca**	**F_1F1**
0	50.00%	33.33%	100.00%	50.00%
1	65.00%	57.14%	88.89%	69.56%
2	55.00%	45.45%	62.50%	52.63%
3	70.00%	68.75%	91.67%	78.57%
4	85.00%	80.00%	100.00%	88.89%
5	80.00%	71.43%	100.00%	83.33%
6	75.00%	70.59%	100.00%	82.76%
7	70.00%	62.50%	100.00%	76.92%
8	60.00%	55.56%	100.00%	71.43%
9	70.00%	72.22%	92.86%	81.25%
avg	68.00%	61.70%	93.59%	73.53%

We use the SVM implementation provided by SVM-light. As we can observe from the two tables, using the stratified 10-fold cross-validation can achieve better average testing performance.

Table 2: SVM results on the stratified 10-fold cross-validation

k	**Acc**	**Prec**	**Reca**	**F_1F1**
0	85.00%	78.57%	100.00%	88.00%
1	40.00%	25.00%	100.00%	40.00%
2	75.00%	75.00%	92.31%	82.76%
3	65.00%	58.82%	100.00%	74.07%
4	65.00%	56.25%	100.00%	72.00%
5	75.00%	76.92%	83.33%	80.00%
6	70.00%	64.71%	100.00%	78.57%
7	70.00%	57.14%	100.00%	72.72%
8	75.00%	78.57%	84.62%	81.48%
9	60.00%	58.82%	90.91%	71.43%
avg	68.00%	62.98%	95.12%	74.10%

Logistic Regression

We use the 50 features as the input of the logistic regression model and predict the label of an input. The results are presented in Table 3 and 4. We can see that the logistic regression model performs worse than the SVM model. We also observe that using the stratified 10-fold cross-validation can achieve better average testing performance.

Table 3: Logistic Regression on the standard 10-fold cross-validation

k	**Acc**	**Prec**	**Reca**	**F_1F1**
0	45.00%	31.25%	100.00%	47.62%
1	75.00%	64.29%	100.00%	78.26%
2	60.00%	50.00%	62.50%	55.56%
3	45.00%	55.56%	41.67%	47.62%
4	70.00%	87.50%	58.33%	70.00%
5	55.00%	57.14%	40.00%	47.06%
6	50.00%	60.00%	50.00%	54.55%
7	50.00%	50.00%	50.00%	50.00%
8	55.00%	57.14%	40.00%	47.06%
9	55.00%	72.73%	57.14%	64.00%
avg	56.00%	58.56%	59.96%	56.17%

Table 4: Logistic Regression on the stratified 10-fold cross-validation

k	**Acc**	**Prec**	**Reca**	**F_1F1**
0	80.00%	76.92%	90.91%	83.33%
1	35.00%	20.00%	75.00%	31.58%
2	60.00%	72.73%	61.54%	66.67%
3	65.00%	63.64%	70.00%	66.67%
4	50.00%	45.45%	55.56%	50.00%
5	85.00%	90.91%	83.33%	86.96%
6	80.00%	76.92%	90.91%	83.33%
7	55.00%	46.15%	75.00%	57.14%
8	65.00%	75.00%	69.23%	72.00%
9	45.00%	50.00%	45.45%	47.62%
avg	62.00%	61.77%	71.69%	64.53%

Random Forest

We use a small random forest model with just 3 trees and a maximal depth of 2. We also enable the bootstrap function. The results are presented in Table 5 and 6. We can see that the random forest model performs better than the logistic regression model on the standard cross-validation. However, the random forest model performs worse than the SVM model on the stratified cross-validation.

Table 5: Random forest on the standard 10-fold cross-validation

k	Acc	Prec	Reca	F_1F1
0	70.00%	33.33%	20.00%	25.00%
1	65.00%	75.00%	33.33%	46.15%
2	45.00%	38.46%	62.50%	47.62%
3	70.00%	68.75%	91.67%	78.57%
4	85.00%	80.00%	100.00%	88.89%
5	60.00%	66.67%	40.00%	50.00%
6	80.00%	75.00%	100.00%	85.71%
7	50.00%	50.00%	30.00%	37.50%
8	70.00%	66.67%	80.00%	72.73%
9	55.00%	69.23%	64.29%	66.67%
avg	65.00%	62.31%	62.18%	59.88%

Table 6: Random forest on the stratified 10-fold cross-validation

k	Acc	Prec	Reca	F_1F1
0	55.00%	75.00%	27.27%	40.00%
1	35.00%	20.00%	75.00%	31.58%
2	45.00%	75.00%	23.08%	35.30%
4	50.00%	33.33%	11.11%	16.66%
5	75.00%	76.92%	83.33%	80.00%
6	75.00%	68.75%	100.00%	81.48%
7	60.00%	50.00%	50.00%	50.00%
8	75.00%	75.00%	92.31%	82.76%
9	55.00%	57.14%	72.73%	64.00%
avg	57.50%	59.01%	53.48%	53.54%

Deep Neural Network

We also implement a simple deep neural network to perform this classification task. This is a 3-layer DNN model, each layer is connected via the sigmoid activation function. We use SGD as the optimizer and set the learning rate as 0.01. The input to this model is normalized to achieve faster training. The results are presented in Table 7 and 8. We can see that the DNN model performs better than both the random forest model and the logistic regression model. Similarly, the results on the stratified cross-validation are slightly better than the results on the standard cross-validation.

However, none of the additional classifiers outperform the original SVM model.

Replication Package

We provide the replication package with the submission. The package contains a README file thoroughly documenting the steps to replicating our results.

Table 7: DNN on the standard 10-fold cross-validation

k	Acc	Prec	Reca	F_1F1
0	50.00%	30.77%	80.00%	44.45%
1	60.00%	53.85%	77.78%	63.64%
2	60.00%	50.00%	62.50%	55.56%
3	60.00%	64.29%	75.00%	69.23%
4	80.00%	83.33%	83.33%	83.33%
5	75.00%	66.67%	100.00%	80.00%
6	75.00%	81.82%	75.00%	78.26%
7	60.00%	58.33%	70.00%	63.63%
8	55.00%	55.56%	50.00%	52.63%
9	50.00%	66.67%	57.14%	61.54%
avg	62.50%	61.13%	73.07%	65.23%

Table 8: DNN on the stratified 10-fold cross-validation

k	Acc	Prec	Reca	F_1F1
0	85.00%	83.33%	90.91%	86.96%
1	35.00%	20.00%	75.00%	31.58%
2	65.00%	75.00%	69.23%	72.00%
3	70.00%	66.67%	80.00%	72.73%
4	50.00%	45.45%	55.56%	50.00%
5	80.00%	83.33%	83.33%	83.33%
6	75.00%	71.43%	90.91%	80.00%
7	55.00%	46.15%	75.00%	57.14%
8	65.00%	80.00%	61.54%	69.57%
9	55.00%	60.00%	54.55%	57.15%
avg	63.50%	63.14%	73.60%	66.05%

References

[1] Cross-validation (statistics), Feb 2022.