Task 1

Confusion matrix: is a standard format for accuracy evaluation, expressed in the matrix form of n-row n-column. Specific evaluation indicators have overall accuracy, mapping accuracy, user accuracy, these accuracy indicators from different sides reflect the accuracy of image classification.

	PREDICTED CLASS		
ACTURAL CLASS		Class=Yes	Class=No
	Class=Yes	a （TP）	b （FN）
	Class=No	c （FP）	d （TN）

Precision metric: biased towards C(Yes|Yes) & C(Yes|No).

$$precision = \frac{TP}{TP + FP} = \frac{a}{a + c}$$

Recall metric: biased towards C(Yes|Yes) & C(No|Yes).

$$recall = \frac{TP}{TP + FN} = \frac{a}{a + b}$$

F-measure metric: a harmonic mean of precision and recall, is biased towards all except C(No|No).

$$f1 = \frac{2 * recall * precision}{recall + precision} = \frac{2 * TP}{2 * TP + FP + FN} = \frac{2a}{2a + b + c}$$

Accuracy metric: Number of correct predictions Total number of predictions.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{a + d}{a + b + c + d}$$

	Precision	Recall	F1	Accuracy
LR model-1	0.9012	0.9124	0.9028	0.9124
LR model-2	0.7908	0.8893	0.8371	0.8893

From the result between LR model-1 and LR model-2, we can see that using normalized numerical features can train mode better.

Task 2

1. Imbalanced data may cause the model misleading.

2. The way to avoid the imbalance issue:

(1) collect more data;

(2) Modify the distribution of training data so that rare class is well-represented in training set.

(3) Cost-based Approaches, introducing different misclassification costs, assigning different weights to classes in training cost function.

	Precision	Recall	F1	Accuracy
Imbalanced data	0.9012	0.9124	0.9028	0.9124
Balanced data	0.9147	0.904	0.9084	0.904

Task 3

Feature selection can Eliminate irrelevant or redundant features to reduce the number of features, improve model accuracy, and reduce runtime.

	Precision	Recall	F1	Accuracy

Original data	0.9012	0.9124	0.9028	0.9124
Partial dataset(k=1)	0.8707	0.895	0.8711	0.895
Partial dataset(k=3)	0.8781	0.898	0.880	0.898
Partial dataset(k=5)	0.8959	0.909	0.8963	0.909

Task 4

Accuracy	LR	Decision tree	SVM	MLP
Imbalanced data	0.9124	0.8917	0.8995	0.9083
Balanced data	0.904	0.8368	0.7393	0.8507

Default parameter settings:

(1) Logisitic Regression:

penalty='l2', dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, class_weight=None, random_state=None, solver='lbfgs', max_iter=100, multi_class='auto', verbose=0, warm_start=False, n_jobs=None, l1_ratio=None

(2) Decision tree:

criterion='gini', splitter='best', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=None, random_state=None, max_leaf_nodes=None, min_impurity_decrease=0.0, class_weight=None, ccp_alpha=0.0

(3) SVM:

C=1.0, kernel='rbf', degree=3, gamma='scale', coef0=0.0, shrinking=True, probability=False, tol=0.001, cache_size=200, class_weight=None, verbose=False, max_iter=- 1, decision_function_shape='ovr', break_ties=False, random_state=None

(4) MLP:

*hidden_layer_sizes=(100), activation='relu', *, solver='adam', alpha=0.0001, batch_size='auto', learning_rate='constant', learning_rate_init=0.001, power_t=0.5, max_iter=200, shuffle=True, random_state=None, tol=0.0001, verbose=False, warm_start=False, momentum=0.9, nesterovs_momentum=True, early_stopping=False, validation_fraction=0.1, beta_1=0.9, beta_2=0.999, epsilon=1e-08, n_iter_no_change=10, max_fun=15000*

Task 5

```
LogisticRegression best score: 0.906
LogisticRegression best parameters:
 {'C': 1, 'class_weight': None, 'dual': False, 'fit_intercept': Tru
e, 'intercept_scaling': 1, 'l1_ratio': None, 'max_iter': 100, 'mult
i_class': 'auto', 'n_jobs': None, 'penalty': 'l2', 'random_state':
 None, 'solver': 'lbfgs', 'tol': 0.0001, 'verbose': 0, 'warm_start
': False}
Decision tree best score: 0.909
Decision tree best parameters:
 {'ccp_alpha': 0.0, 'class_weight': None, 'criterion': 'gini', 'max
_depth': 30, 'max_features': None, 'max_leaf_nodes': None, 'min_im
purity_decrease': 0.0, 'min_impurity_split': None, 'min_samples_le
af': 20, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0,
'random_state': None, 'splitter': 'best'}
```

Best model on test data:

```
Precision Score:  0.9047905206767384
Recall Score:  0.9106
F1 Score:  0.907097021076788
Accuracy Score:  0.9106
```