MUHAMMAD SYAFIQ BIN AZHARI (289869)

School Of Computing,

UUM College Of Arts and Sciences

SKIH3013 PATTERN RECOGNITION AND ANALYSIS

Semester A231

SKIH3013 (PATTERN RECOGNITION & ANALYSIS) – A231 Assignment (IV)

Title:
Loan Approval (Manipulation of Hyperparameters)

Prepared by:

Name	Matric No
Muhammad Syafiq bin Azhari	289869

Submission Date: 27th December 2023

Lecturer's Name: ASSOCIATE PROF. DR. AZIZI AB AZIZ

MUHAMMAD SYAFIQ BIN AZHARI (289869)

Question:

Tasks – You are required to use Python programming language (with Scikit-learn, and Seaborn libraries) :

Based on the dataset *LoanApprovalPrediction*.csv and Python codes (file: *SKIH3013-Assgn04-LoanApprovalApplication.py*), you are required to do the following:

1. Explain the THREE (3) pros and cons of each classifier.
2. Modify the given Python codes to optimize all pattern recognition classifiers (if any) hyperparameters by using the Grid Search Hyperparameter tuning. As a basis, you are required to tune each classifier based on these initialised hyperparameters.
    - K-Nearest Neighbours

      ```
      params = { 'n_neighbors' : [3, 5,7,9,11,13,15],
                 'weights' : ['uniform','distance'],
                 'metric' : ['minkowski','euclidean','manhattan']}
      ```

    - Decision Tree

      ```
      params = { 'max_depth': [2, 3, 5, 10, 20],
        'min_samples_leaf': [5, 10, 20, 50, 100],
        'criterion': ["gini", "entropy"]}
      ```

    - Random Forest

      ```
      params = {
        'n_estimators': [25, 50, 100, 150],
        'max_features': ['sqrt', 'log2', None],
        'max_depth': [3, 6, 9],
        'max_leaf_nodes': [3, 6, 9], }
      ```

    - Logistic Regression

      ```
      param = {
        'penalty' : ['l1', 'l2', 'none'],
        'solver' : ['lbfgs','newton-cg'],
        'max_iter' : [100, 1000, 300],  }
      ```

    - Support Vector Machine

      ```
      param = {'C': [0.1, 1, 10, 100],
                'gamma': [1, 0.5, 0.1, 0.001],
                'kernel': ['rbf', 'linear']}
      ```

    - Multilayer Perceptron (Backpropagation)

      ```
      param = {
        'hidden_layer_sizes': [(150,100,50), (120,80,40), (100,50,30)],
        'max_iter': [100, 500, 1000],
        'activation': ['tanh', 'relu'],
        'solver': ['sgd', 'adam'],
        'alpha': [0.0001, 0.05],
        'learning_rate': ['constant','adaptive']}
      ```

3. Plot the ROC/AUC graph to show the classification results for each classifier based on tuned hyperparameters.
4. Decide and describe the best classifier.
5. Explain the advantages and disadvantages of using a majority voting approach for this case study.
6. Get the classification results for these new datasets based on the majority voting approach (using the optimized hyperparameter tuning):

   new_case = < Gender, Married, Dependents, Education, Self Employed, Applicant Income Co-applicant Income, Loan Amount, Loan Amount Term, Credit History, Property Area, Loan Status >

    - A = *< Male, Yes, 2, Graduate, No, 11417, 1126, 225, 360, 1, Urban, ?>*
    - B = *< Male, Yes, 3, Not Graduate, Yes,5703,0,130, 360,1, Rural, ?>*
    - C = *< Female, Yes,0, Graduate,No,4333,2451,110,360,1, Urban, ?>*

MUHAMMAD SYAFIQ BIN AZHARI (289869)

Task 1:

Classifier	Pros	Cons
K-Nearest Neighbour	● Simple distance-based algorithm ● Can handle large datasets ● Does not make assumptions	● Is highly sensitive to outliers ● Suffers the curse of dimensionality ● Requires to find optimal value of K for optimal classification
Decision Tree	● Not affected by missing values ● Easily explainable (can see the rules) ● Can handle both numerical and categorical values	● Very easy to fall under overfitting when the depth of the tree is not controlled ● Becomes increasingly complex as the depth increases ● Small changes can make the structure of the decision tree to change dramatically causing instability
Random Forest	● Very quick to classify once it has passed the training stage ● Takes less computational power when used to classify ● Like decision tree (it literally is a bunch of decision trees), it can handle both numerical and categorical values	● Takes more time for training ● Takes more computational power than that of decision tree during training ● Hard to interpret and explain
Logistic Regression	● Incredibly easy to use ● Incredibly fast and efficient in classifying unknown records ● It makes no assumptions about distributions of classes in feature space.	● Could lead to potential overfitting if number of records are less than number of features ● It can only handle linear problems ● Sensitive to outliers
Support Vector Machine (SVM)	● Handles increasing dimensionality well ● Works well with unstructured or semi-structured data ● Has strong generalisation power	● Takes very long to train ● Does not handle large data sets well ● Difficult to find the best hyperparameters

MUHAMMAD SYAFIQ BIN AZHARI (289869)

Multilayer Perceptron (MLP)	• Able to solve complex nonlinear problems • Handles large amounts of data sets • Very quick classification after training	• It takes a long time to train • Less effective with overlapping and noisy datasets • Difficult to find the best hyperparameters (which then increases training time)

MUHAMMAD SYAFIQ BIN AZHARI (289869)

Task 2

Test #1:

```
----- Best Hyperparameters -----
--------KNN--------
{'metric': 'minkowski', 'n_neighbors': 9, 'weights': 'uniform'}
--------DT--------
{'criterion': 'gini', 'max_depth': 2, 'min_samples_leaf': 10}
--------RF--------
{'max_depth': 6, 'max_features': 'sqrt', 'max_leaf_nodes': 9, 'n_estimators': 25}
--------LR--------
{'max_iter': 100, 'penalty': 'l2', 'solver': 'lbfgs'}
--------SVC--------
{'C': 0.1, 'gamma': 1, 'kernel': 'linear'}
--------MLP--------
{'activation': 'relu', 'alpha': 0.0001, 'hidden_layer_sizes': (150, 100, 50), 'learning_rate': 'adaptive', 'max_iter': 500,
'solver': 'sgd'}
```

```
--------------AFTER HYPER PARAMETER TUNING RESULTS-----

--------------KNN--------------------
              precision    recall  f1-score   support

           0       0.89      0.52      0.65        33
           1       0.84      0.98      0.90        87

    accuracy                           0.85       120
   macro avg       0.87      0.75      0.78       120
weighted avg       0.86      0.85      0.84       120


Confusion matrix using KNN:
[[17 16]
 [ 2 85]]
```

MUHAMMAD SYAFIQ BIN AZHARI (289869)

```
--------------DT-------------------
              precision    recall  f1-score   support

           0       0.94      0.52      0.67        33
           1       0.84      0.99      0.91        87

    accuracy                           0.86       120
   macro avg       0.89      0.75      0.79       120
weighted avg       0.87      0.86      0.84       120


Confusion matrix using DT:
[[17 16]
 [ 1 86]]

--------------RF-------------------
              precision    recall  f1-score   support

           0       0.94      0.52      0.67        33
           1       0.84      0.99      0.91        87

    accuracy                           0.86       120
   macro avg       0.89      0.75      0.79       120
weighted avg       0.87      0.86      0.84       120


Confusion matrix using RF:
[[17 16]
 [ 1 86]]
```

MUHAMMAD SYAFIQ BIN AZHARI (289869)

```
--------------LR-------------------
              precision    recall  f1-score   support

           0       0.94      0.52      0.67        33
           1       0.84      0.99      0.91        87

    accuracy                           0.86       120
   macro avg       0.89      0.75      0.79       120
weighted avg       0.87      0.86      0.84       120


Confusion matrix using LR:
[[17 16]
 [ 1 86]]

--------------SVC-------------------
              precision    recall  f1-score   support

           0       0.94      0.52      0.67        33
           1       0.84      0.99      0.91        87

    accuracy                           0.86       120
   macro avg       0.89      0.75      0.79       120
weighted avg       0.87      0.86      0.84       120


Confusion matrix using SVC:
[[17 16]
 [ 1 86]]

--------------MLP-------------------
              precision    recall  f1-score   support

           0       0.94      0.52      0.67        33
           1       0.84      0.99      0.91        87

    accuracy                           0.86       120
   macro avg       0.89      0.75      0.79       120
weighted avg       0.87      0.86      0.84       120


Confusion matrix using MLP:
[[17 16]
 [ 1 86]]
```

MUHAMMAD SYAFIQ BIN AZHARI (289869)

Test #2

```
----- Best Hyperparameters -----
--------KNN--------
{'metric': 'minkowski', 'n_neighbors': 9, 'weights': 'uniform'}
--------DT--------
{'criterion': 'gini', 'max_depth': 2, 'min_samples_leaf': 10}
--------RF--------
{'max_depth': 6, 'max_features': 'sqrt', 'max_leaf_nodes': 9, 'n_estimators': 25}
--------LR--------
{'max_iter': 100, 'penalty': 'l2', 'solver': 'lbfgs'}
--------SVC--------
{'C': 0.1, 'gamma': 1, 'kernel': 'linear'}
--------MLP--------
{'activation': 'relu', 'alpha': 0.0001, 'hidden_layer_sizes': (120, 80, 40), 'learning_rate': 'adaptive', 'max_iter': 500, 's
olver': 'sgd'}
```

```
--------------AFTER HYPER PARAMETER TUNING RESULTS-----

--------------KNN--------------------
              precision    recall  f1-score   support

           0       0.89      0.52      0.65        33
           1       0.84      0.98      0.90        87

    accuracy                           0.85       120
   macro avg       0.87      0.75      0.78       120
weighted avg       0.86      0.85      0.84       120


Confusion matrix using KNN:
[[17 16]
 [ 2 85]]

--------------DT--------------------
              precision    recall  f1-score   support

           0       0.94      0.52      0.67        33
           1       0.84      0.99      0.91        87

    accuracy                           0.86       120
   macro avg       0.89      0.75      0.79       120
weighted avg       0.87      0.86      0.84       120


Confusion matrix using DT:
[[17 16]
 [ 1 86]]

--------------RF--------------------
              precision    recall  f1-score   support

           0       0.94      0.52      0.67        33
           1       0.84      0.99      0.91        87

    accuracy                           0.86       120
   macro avg       0.89      0.75      0.79       120
weighted avg       0.87      0.86      0.84       120


Confusion matrix using RF:
[[17 16]
 [ 1 86]]
```

MUHAMMAD SYAFIQ BIN AZHARI (289869)

```
--------------LR-------------------
              precision    recall  f1-score   support

           0       0.94      0.52      0.67        33
           1       0.84      0.99      0.91        87

    accuracy                           0.86       120
   macro avg       0.89      0.75      0.79       120
weighted avg       0.87      0.86      0.84       120


Confusion matrix using LR:
[[17 16]
 [ 1 86]]

--------------SVC-------------------
              precision    recall  f1-score   support

           0       0.94      0.52      0.67        33
           1       0.84      0.99      0.91        87

    accuracy                           0.86       120
   macro avg       0.89      0.75      0.79       120
weighted avg       0.87      0.86      0.84       120


Confusion matrix using SVC:
[[17 16]
 [ 1 86]]

--------------MLP-------------------
              precision    recall  f1-score   support

           0       0.94      0.52      0.67        33
           1       0.84      0.99      0.91        87

    accuracy                           0.86       120
   macro avg       0.89      0.75      0.79       120
weighted avg       0.87      0.86      0.84       120


Confusion matrix using MLP:
[[17 16]
 [ 1 86]]
```

MUHAMMAD SYAFIQ BIN AZHARI (289869)

Test #3

```
----- Best Hyperparameters -----
--------KNN--------
{'metric': 'minkowski', 'n_neighbors': 9, 'weights': 'uniform'}
--------DT--------
{'criterion': 'gini', 'max_depth': 2, 'min_samples_leaf': 10}
--------RF--------
{'max_depth': 3, 'max_features': 'sqrt', 'max_leaf_nodes': 3, 'n_estimators': 25}
--------LR--------
{'max_iter': 100, 'penalty': 'l2', 'solver': 'lbfgs'}
--------SVC--------
{'C': 0.1, 'gamma': 1, 'kernel': 'linear'}
--------MLP--------
{'activation': 'relu', 'alpha': 0.05, 'hidden_layer_sizes': (150, 100, 50), 'learning_rate': 'constant', 'max_iter': 500, 'so
lver': 'sgd'}
```

```
--------------AFTER HYPER PARAMETER TUNING RESULTS----

--------------KNN--------------------
              precision    recall  f1-score   support

           0       0.89      0.52      0.65        33
           1       0.84      0.98      0.90        87

    accuracy                           0.85       120
   macro avg       0.87      0.75      0.78       120
weighted avg       0.86      0.85      0.84       120


Confusion matrix using KNN:
[[17 16]
 [ 2 85]]

--------------DT--------------------
              precision    recall  f1-score   support

           0       0.94      0.52      0.67        33
           1       0.84      0.99      0.91        87

    accuracy                           0.86       120
   macro avg       0.89      0.75      0.79       120
weighted avg       0.87      0.86      0.84       120


Confusion matrix using DT:
[[17 16]
 [ 1 86]]

--------------RF--------------------
              precision    recall  f1-score   support

           0       0.94      0.52      0.67        33
           1       0.84      0.99      0.91        87

    accuracy                           0.86       120
   macro avg       0.89      0.75      0.79       120
weighted avg       0.87      0.86      0.84       120


Confusion matrix using RF:
[[17 16]
 [ 1 86]]
```

MUHAMMAD SYAFIQ BIN AZHARI (289869)

```
--------------LR-------------------
              precision    recall  f1-score   support

           0       0.94      0.52      0.67        33
           1       0.84      0.99      0.91        87

    accuracy                           0.86       120
   macro avg       0.89      0.75      0.79       120
weighted avg       0.87      0.86      0.84       120


Confusion matrix using LR:
[[17 16]
 [ 1 86]]

--------------SVC-------------------
              precision    recall  f1-score   support

           0       0.94      0.52      0.67        33
           1       0.84      0.99      0.91        87

    accuracy                           0.86       120
   macro avg       0.89      0.75      0.79       120
weighted avg       0.87      0.86      0.84       120


Confusion matrix using SVC:
[[17 16]
 [ 1 86]]

--------------MLP-------------------
              precision    recall  f1-score   support

           0       0.94      0.52      0.67        33
           1       0.84      0.99      0.91        87

    accuracy                           0.86       120
   macro avg       0.89      0.75      0.79       120
weighted avg       0.87      0.86      0.84       120


Confusion matrix using MLP:
[[17 16]
 [ 1 86]]
```

MUHAMMAD SYAFIQ BIN AZHARI (289869)

Task 3

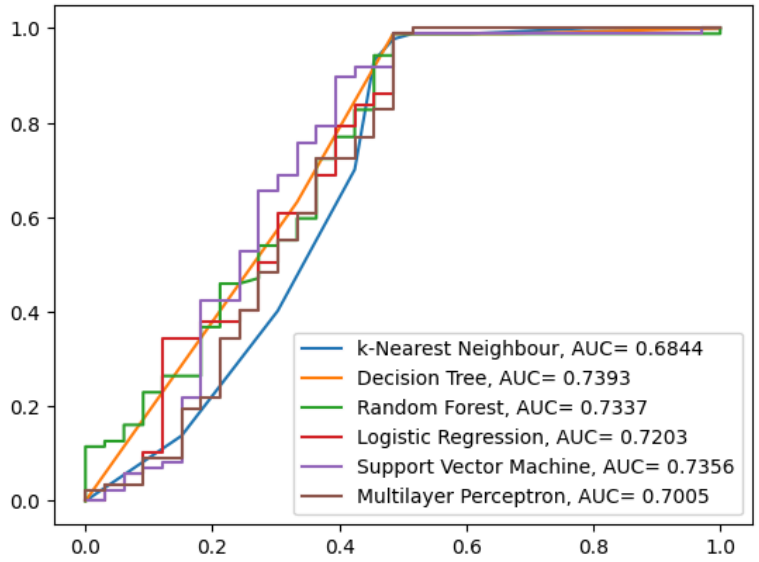

Figure 3.1: ROC Graph of Syafiq's Classifiers (Test #1)

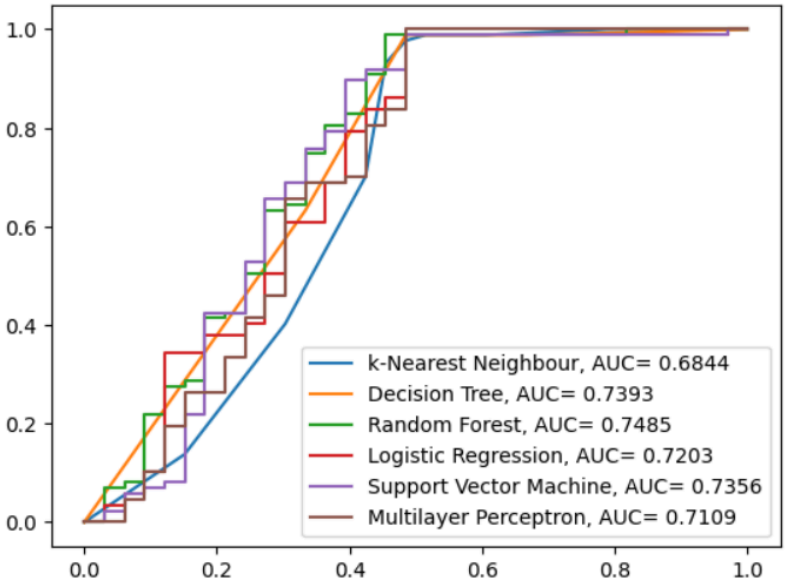

Figure 3.2: ROC Graph of Syafiq's Classifiers (Test #2)

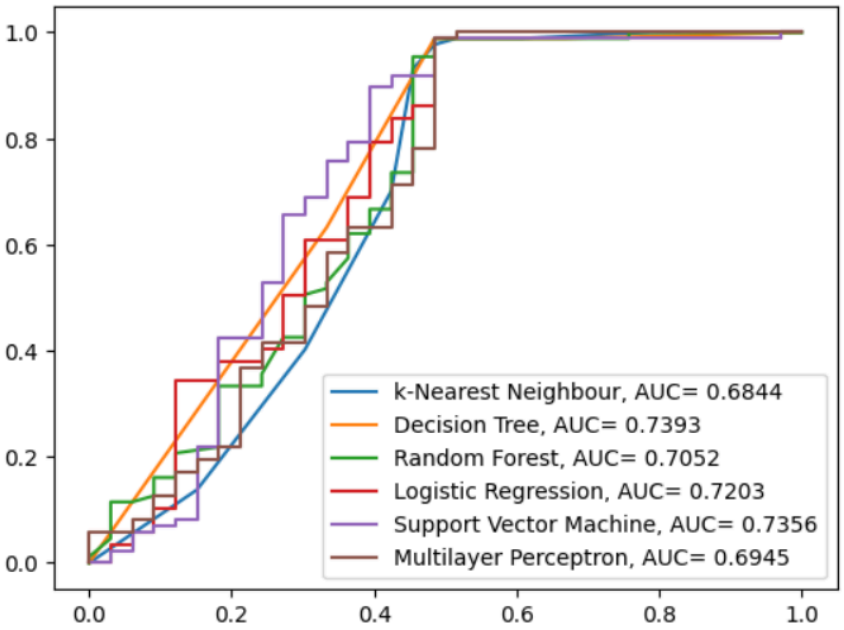

Figure 3.3: ROC Graph of Syafiq's Classifiers (Test #3)

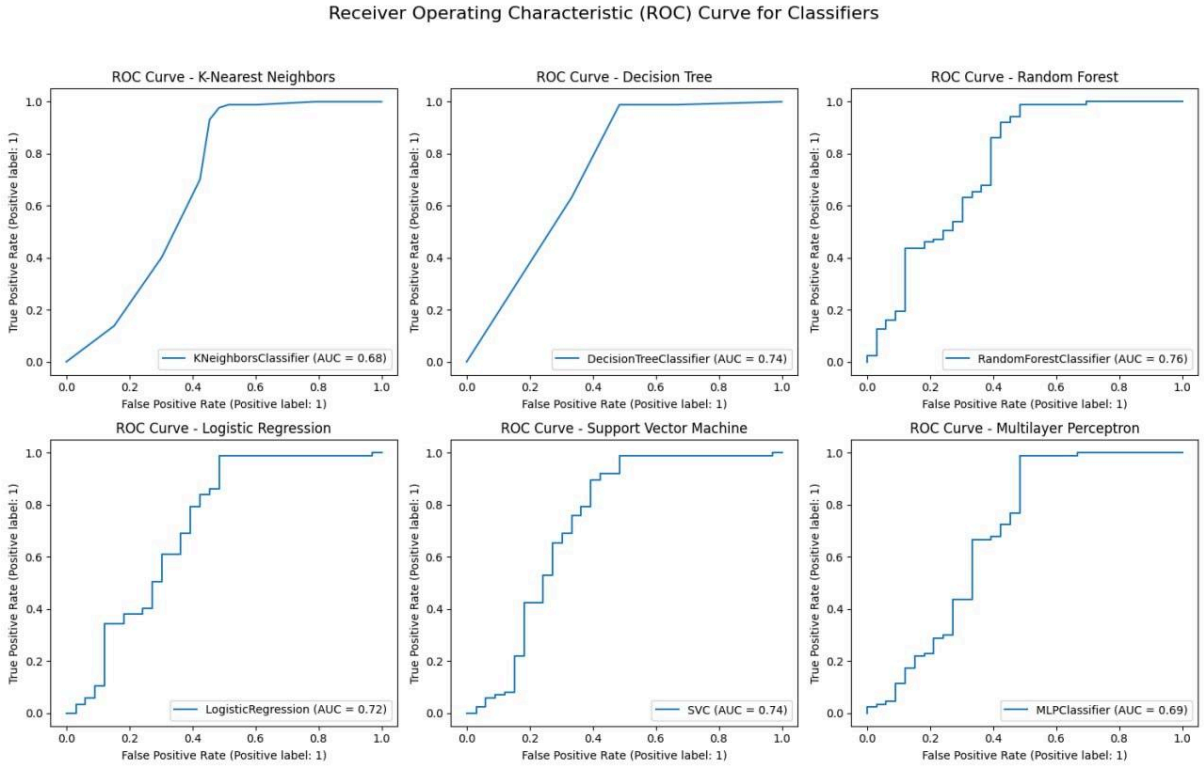

Figure 3.3: ROC Graphs of Suchira's Test

MUHAMMAD SYAFIQ BIN AZHARI (289869)

Task 4

To properly get the best result, this project was run 3 times to see if there were any changes. As seen above, quite significant changes were seen as in Test #1, the classifier with the best AUC score was Decision Tree with a score of 0.7393 with Random Forest coming in second with 0.7337. However in Test #2, Random Forest obtained a new AUC highscore with 0.7485 with Decision Tree coming in second at 0.7393. On the last (Test #3), the results are worse than the first two with the highest AUC score 0.7393 by Decision Tree. To compare my results with a peer, we can take a look at Suchira's (288316) test run where the AUC score achieved by her Random Forest classifier was 0.76.

Thus, in terms of which is the best classifier, I would have to choose Random Forest as it has both the highest AUC score in my test and Suchira's test has an even higher Random Forest AUC score than my highest.

My best Random Forest Hyperparameters:

```
--------RF--------
{'max_depth': 6, 'max_features': 'sqrt', 'max_leaf_nodes': 9, 'n_estimators': 25}
```

MUHAMMAD SYAFIQ BIN AZHARI (289869)

Task 5

In this case study, the advantage of having the majority voting system is that each classifier would be able to classify different things much better than other classifiers. For instance, hypothetically the Decision Tree classifier could better classify loans that should be approved and Support Vector Machine (SVM) would be better at classifying loans that should be ejected. The majority vote in a sense would be able to take advantage of the individual strengths of each classifier with a few minor setbacks.

The setbacks are that if the majority of the classifiers are good at classifying loans that should be approved, the ability of the whole system to classify a loan that should be rejected becomes questionable if not dangerous. This bias is caused by the system taking the votes as it is without any prior or further considerations done.

A way to overcome this is to give weights to the votes. In other words, each classifier would be given a different priority, meaning a different classifier has a more powerful vote than others. The strength of the votes can be attributed to each classifier's accuracy score or any other similar and suitable metric.

MUHAMMAD SYAFIQ BIN AZHARI (289869)

Task 6

Case A:

Test #1:

```
new classification for: [    1     0     2     0     0 11417  1126   225   360     1     1]
Classification: k-Nearest Neighbour --> [1]
Classification: Decision Tree --> [1]
Classification: Random Forest --> [1]
Classification: Logistic Regression --> [1]
Classification: Support Vector Machine --> [1]
Classification: Multilayer Perceptron --> [1]

 +++++++++++++++++ Predicting a new case ++++++++++++++++++++

new classification for:
 [    1     0     2     0     0 11417  1126   225   360     1     1]

--------------------- Final Result ---------------------------------
The applicant loan will be approved with overall 100.00% vote
```

Test #2

```
new classification for: [    1     0     2     0     0 11417  1126   225   360     1     1]
Classification: k-Nearest Neighbour --> [1]
Classification: Decision Tree --> [1]
Classification: Random Forest --> [1]
Classification: Logistic Regression --> [1]
Classification: Support Vector Machine --> [1]
Classification: Multilayer Perceptron --> [1]

 +++++++++++++++++ Predicting a new case ++++++++++++++++++++

new classification for:
 [    1     0     2     0     0 11417  1126   225   360     1     1]

--------------------- Final Result ---------------------------------
The applicant loan will be approved with overall 100.00% vote
```

Test #3

MUHAMMAD SYAFIQ BIN AZHARI (289869)

Case B:

Test #1:

```
new classification for: [   1    0    3    1    1 5703    0  130  360    1    0]
Classification: k-Nearest Neighbour --> [1]
Classification: Decision Tree --> [1]
Classification: Random Forest --> [1]
Classification: Logistic Regression --> [1]
Classification: Support Vector Machine --> [1]
Classification: Multilayer Perceptron --> [1]

 ++++++++++++++++++ Predicting a new case ++++++++++++++++++++++

new classification for:
 [   1    0    3    1    1 5703    0  130  360    1    0]

--------------------- Final Result --------------------------------
The applicant loan will be approved with overall 100.00% vote
```

Test #2

```
new classification for: [   1    0    3    1    1 5703    0  130  360    1    0]
Classification: k-Nearest Neighbour --> [1]
Classification: Decision Tree --> [1]
Classification: Random Forest --> [1]
Classification: Logistic Regression --> [1]
Classification: Support Vector Machine --> [1]
Classification: Multilayer Perceptron --> [1]

 ++++++++++++++++++ Predicting a new case ++++++++++++++++++++++

new classification for:
 [   1    0    3    1    1 5703    0  130  360    1    0]

--------------------- Final Result --------------------------------
The applicant loan will be approved with overall 100.00% vote
```

Test #3

MUHAMMAD SYAFIQ BIN AZHARI (289869)

Case C:

Test #1

```
new classification for: [   0    0    0    0    0 4333 2451  110  360    1    1]
Classification: k-Nearest Neighbour --> [1]
Classification: Decision Tree --> [1]
Classification: Random Forest --> [1]
Classification: Logistic Regression --> [1]
Classification: Support Vector Machine --> [1]
Classification: Multilayer Perceptron --> [1]

 ++++++++++++++++++ Predicting a new case ++++++++++++++++++++++

new classification for:
 [   0    0    0    0    0 4333 2451  110  360    1    1]

--------------------- Final Result --------------------------------
The applicant loan will be approved with overall 100.00% vote
```

Test #2

```
new classification for: [   0    0    0    0    0 4333 2451  110  360    1    1]
Classification: k-Nearest Neighbour --> [1]
Classification: Decision Tree --> [1]
Classification: Random Forest --> [1]
Classification: Logistic Regression --> [1]
Classification: Support Vector Machine --> [1]
Classification: Multilayer Perceptron --> [1]

 ++++++++++++++++++ Predicting a new case ++++++++++++++++++++++

new classification for:
 [   0    0    0    0    0 4333 2451  110  360    1    1]

--------------------- Final Result --------------------------------
The applicant loan will be approved with overall 100.00% vote
```

Test #3

In all three cases, every classifier has voted to approve the loan request resulting in a 100.00% vote in all 3 cases.