Assignment 3: Feature Extraction and Classification Techniques

Date	FUND01	FUND02	FUND03	..	FUND200
15/01/1993	0.0100	0.0088	0.0367	..	0.0357
15/02/1993	0.0161	0.0146	0.0185	..	0.0232
15/03/1993	0.0732	0.0629	0.0018	..	0.0281
..	..	..	..	..	..
15/12/2022	-0.0348	-0.0235	-0.0392	..	-0.0727

Table 1: Fund Performance

Fund Category	Explanation
INT_VAL	International value-oriented fund
INT_ACT	Actively managed international fund
US_VAL	U.S. value-oriented fund
US_ACT	Actively managed U.S. fund

Table 2: Fund Categories/Labels

Date	EURMKT	EURSIZ	EURVAL	..	USDMOM
15/01/1993	0.006	0.032	0.029	..	0.049
15/02/1993	0.011	-0.007	0.017	..	0.027
15/03/1993	0.054	0.014	0.008	..	0.041
..	..	..	..	..	..
15/12/2022	-0.010	0.014	0.025	..	0.040

Table 3: Market Data

FUND	Mean	Vol	Sharpe	Corr_EURMKT	..
FUND01	0.003	0.036	0.078	0.650	..
FUND02	0.003	0.024	0.131	0.474	..
FUND03	0.005	0.029	0.166	0.559	..
..	..	..	..	..	..
FUND200	0.007	0.052	0.141	0.777	..

Table 4: Feature Matrix (200 X 39)

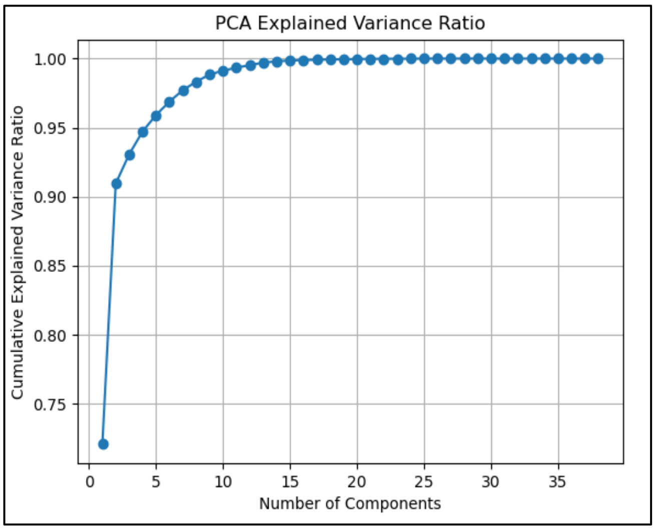

Chart 1: Dimensionality Reduction using PCA

Part A: Feature Extraction

Introduction: The report presents an analysis of feature extraction applied to a dataset comprising 200 funds categorized into four distinct categories. It evaluates the performance of three classification algorithms: Random Forest Classifier, Support Vector Machine, and Multi-Layer Perceptron Classifier, in accurately categorizing these funds into their respective labels. Additionally, the report provides a comparative assessment of the predictive scores and confusion matrices generated by each classification model.

About the Data: The Data itself Consists of Three tabs:

1. ***Fund Performance:*** This section comprises historical return time series data spanning from 1993 to 2022 for a total of 200 distinct funds. These return time series serve as the foundation for generating features used in subsequent analysis.
2. ***Labels:*** Within this segment, a comprehensive list of the 200 funds is provided alongside their respective labels. These funds are categorized into four primary classes:
    a. INT_VAL
    b. INT_ACT
    c. US_VAL
    d. US_ACT
3. ***Market Data:*** This component comprises return series data for Market-Fund types, offering insights into how different fund types perform across diverse market conditions. Combined with the fund performance data, this information is instrumental in feature generation for classification purposes.

Problem Statement:

Given the return series of fund-performance, labels of the funds, and the additional Market data, the objective is to extract relevant features that would help us in classifying these funds with the highest accuracy. Classification algorithm used here are:

- Random Forest Classifier
- Support Vector Machine
- Multi-Layer Perceptron Classifier.

Features Extracted:

A total of 39 features were extracted from the fund performance time series and Market Data tab. Notable features included the mean and standard deviation of the fund return series, from which the Sharpe ratio was derived and incorporated as a classification feature. Additionally, correlation, covariance, and linear regression slope values were computed for each fund against 12 market indicators such as EURMKT, EURSIZ, and JPYMKT. This process resulted in a 200x39 feature matrix, which was subsequently scaled using the minmax scaler function to ensure uniform scaling across all features. Furthermore, Principal Component Analysis (PCA) was employed to reduce dimensionality. Chart 1 illustrates the creation of 15 principal components from the 39 features, which were utilized to train the three classifier models. By utilizing only 15 principal components instead of all 39 features, computational resources were conserved, training time was minimized, and overall model accuracy was optimized.

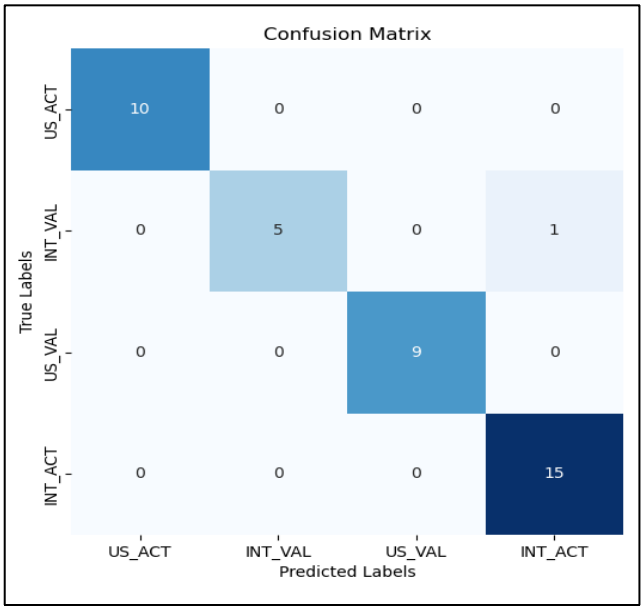

Chart 2: Random Forest

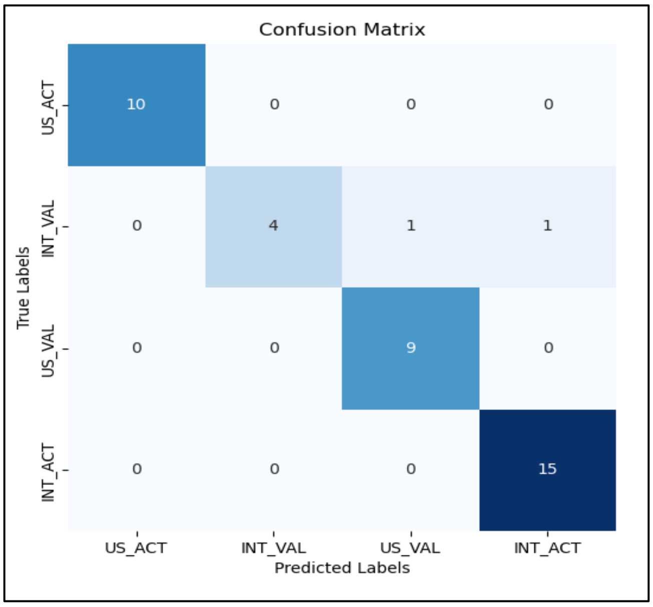

Chart 3: SVM

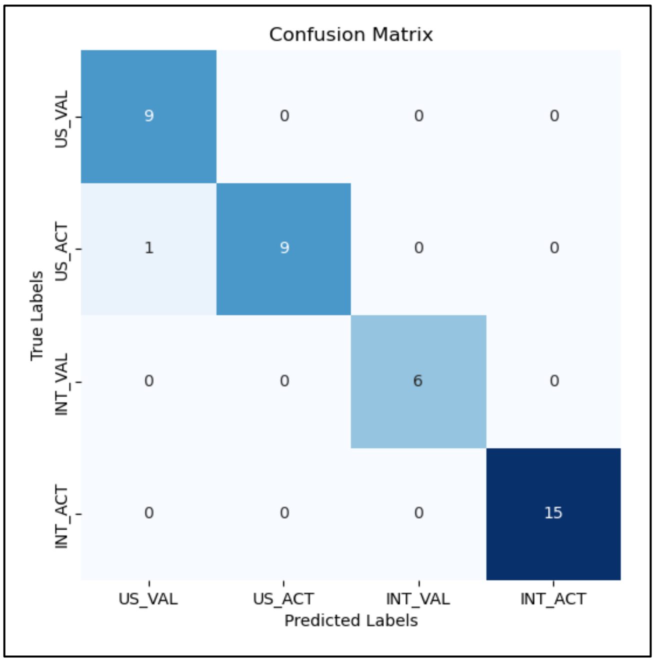

Chart 4: MLP Classifier

Part B: Classification Techniques

Since the dataset provided is small (200 funds) the output from the various classifiers does not largely vary. With MLP classifier performing the best followed by Random Forest and SVM which were not lagging far behind. The detailed performance metrics can be seen in table 5.

Classifier	Performance Metric		
	Accuracy	Precision	F1-score
Random Forest	0.975	0.976	0.974
SVM	0.950	0.954	0.946
MLP	0.975	0.977	0.975

Table 5: Model Performance Metrics

Random Forest Classifier:

Random Forest Classifier is an ensemble learning method that constructs multiple decision trees trained on random subsets of data and features. It aggregates the predictions of these trees through a voting mechanism to make final predictions, resulting in improved accuracy and robustness compared to individual trees.

Chart 2 shows us the confusion matrix for the Random Forest classifier. This matrix depicts the performance of a classifier in a multiclass classification task with four classes. Each row represents the true class, while each column represents the predicted class. The diagonal elements indicate the number of correctly classified instances for each class, while off-diagonal elements represent misclassifications. In this specific case, all classes except one were predicted correctly, with a single instance from the INT_VAL class being misclassified as the INT_ACT class. Overall, the classifier demonstrates strong performance, accurately classifying most instances.

SVM Classifier:

SVM works by finding the optimal hyperplane that separates data points of different classes with the maximum margin. SVM can handle both linear and non-linear classification problems efficiently and is known for its robustness and effectiveness, especially in high-dimensional spaces. SVM achieves an accuracy of 95%, slightly below that of Random Forest. Chart 3 depicts the confusion matrix of the SVM classifier, indicating two incorrect predictions while accurately predicting the remaining observations.

MLP Classifier:

The Multilayer Perceptron (MLP) is a versatile neural network architecture utilized for classification tasks, featuring layers of interconnected neurons with nonlinear activation functions. Through iterative weight adjustments and backpropagation, MLPs efficiently learn to classify input data into different categories, making them valuable tools for diverse classification problems. MLP achieves an accuracy of 97.5%, comparable to that of Random Forest. Chart 4 illustrates the confusion matrix of the SVM classifier, showing one wrong prediction alongside accurately predicted remaining observations.

Part B: Classification Techniques

Feature	Random Forest	SVM	MLP
Advantages	❖ Robust to overfitting. ❖ Handles high-dimensional data. ❖ Provides feature importance.	❖ Effective in high-dimensional spaces. ❖ Versatile kernel functions ❖ Robust to outliers.	❖ Ability to learn complex patterns. ❖ Flexibility in architecture ❖ Handles non-linear decision. boundaries.
Disadvantages	❖ Computationally expensive. ❖ Prone to overfitting.	❖ Computationally intensive. ❖ Sensitivity to hyperparameters. ❖ Limited interpretability.	❖ Prone to overfitting. ❖ Hyperparameter tuning required. ❖ Slow training.
Effectiveness	Random Forests are effective for both classification and regression tasks, offering high accuracy by aggregating predictions from multiple decision trees. They are robust to overfitting and can handle high-dimensional data well.	SVMs are effective in separating data points with a clear margin, especially in high-dimensional spaces. They work well for both linear and non-linear classification tasks and are known for their ability to generalize well to unseen data.	MLPs are versatile and can capture complex patterns in data through multiple hidden layers. They are effective for non-linear classification tasks but are prone to overfitting, especially with large networks and insufficient training data.
Cost function	Typically uses Gini impurity or entropy as the cost function for splitting nodes in decision trees.	Uses hinge loss as the cost function to maximize the margin between support vectors.	Utilizes various cost functions such as cross-entropy loss for classification tasks and mean squared error for regression tasks.
Training time	Generally faster to train compared to SVM and MLP, especially for large datasets.	Training time can be significant, particularly for large datasets, due to the optimization process involving support vectors.	Training time can be slow, especially for large networks and datasets, due to the iterative optimization process and backpropagation.
Interpretability	Provides feature importance insights but may result in complex models with limited interpretability.	Often results in complex models with limited interpretability, especially with non-linear kernels.	Can result in complex models with limited interpretability, especially with multiple hidden layers and neurons.