**COS30018 - Option B - Task 2: Data processing 1**

**Student Name: Nguyen Duc Le Nguyen**

**Id: 104224493**

**Summary of Effort**

This report details the development of a function to load and process stock market data with various features, including handling NaN values, splitting data into train/test sets, scaling features, and saving/loading data locally.

**Code Breakdown and Explanation**

Below is a detailed explanation of the less straightforward lines of code within the function.

A screen shot of a computer

Description automatically generated

**os**: Provides a way of using operating system-dependent functionality, such as reading or writing to the filesystem.

**pandas**: A powerful data manipulation library.

**yfinance**: A library to fetch financial data from Yahoo Finance.

**sklearn.model\_selection.train\_test\_split**: A utility function to split data into train and test sets.

**sklearn.preprocessing.StandardScaler and MinMaxScaler:** Tools for feature scaling.

**joblib:** A library for saving and loading Python objects.

A screen shot of a computer code

Description automatically generated

This function initializes with several parameters, allowing flexibility in data loading, processing, and saving.

A screen shot of a computer code

Description automatically generated

Checks if data should be loaded from a local file. If not, it downloads the data from Yahoo Finance and saves it if required.

A computer screen shot of a computer code

Description automatically generated

Handles NaN values by either dropping them or filling them. Forward fill (ffill) and backward fill (bfill) ensure no NaN values remain.

A black screen with green and orange text

Description automatically generated

Splits the dataframe into features (X) and target (y). Here, 'Adj Close' is assumed to be the target variable.

A screen shot of a computer code

Description automatically generated

Depending on the split\_method, the data is split either randomly according to a specified ratio or by a specific date.

A screen shot of a computer program

Description automatically generated

If scaling is requested, the function applies either StandardScaler or MinMaxScaler to the feature columns. It also saves the scaler if specified.



Returns the processed data splits and the scaler (if applied).

This report covered the key lines of the load\_and\_process\_data function, explaining each part to ensure clarity. Further inquiries about any specific line of code are welcome.