**Stock Movement Prediction Using Market Features and Unsupervised Learning**

Professor Mohammad Toutiaee  
CS6220: Data Mining — Summer 2025

**Group Members**:

Benjamin Northrop (MSCS, 2025)

Carlos Semeho Edorh (MSCS, 2025)

Zachary Coates (MSCS, 2025)

**Description of the Problem**  
 This project explores whether engineered market features can help predict short-term stock movements within the S&P 500. Specifically, can supervised learning techniques reveal latent structure or patterns in historical stock data that inform next-day market direction?

To answer this, we will construct a dataset of financial features such as daily, weekly, and monthly returns, volatility, earnings data, and analyst metrics. We will apply Principal Component Analysis (PCA) to reduce dimensionality, followed by clustering (e.g., K-Means, Agglomerative) to group stocks by behavioral similarity. We also plan to experiment with a simple neural network to predict binary price movement (up or down) for the following day.

**Summary of the Data**  
 We will use the yfinance API to obtain stock-level data for all S&P 500 constituents. Each row will represent one ticker, and columns will include engineered features including but not limited to: daily percentage change (30 days), weekly percentage change (6 months), monthly percentage change (1 year), Earnings Per Share results (as available), Relative Strength Index, Earnings per Share.

We will evaluate the feature set for correlation and completeness. Some tickers may have missing data, such as missing earning reports, which we will handle through imputation or row removal. Features will be standardized to ensure comparability across scales.

**Methods**  
Dimensionality reduction will be performed using PCA, and visualizations will be generated from the top components. Clustering will be applied using both K-Means (with elbow method for k) and Agglomerative Clustering. We will evaluate cluster quality using silhouette scores and Davies-Bouldin Index.

For supervised learning, we will define a binary target based on whether each stock’s price increases or decreases the next day. A feedforward neural network (MLP) will be implemented in Keras and evaluated using accuracy, precision, recall, and F1-score. All code will be run in a reproducible Jupyter notebook using Python libraries such as Pandas, Scikit-learn, and Keras.

A diagram of a computer generated diagram

AI-generated content may be incorrect.

**Preliminary Results**  
We have implemented a prototype notebook to test data collection from yfinance, confirming the feasibility of pulling time series and fundamental data at scale. PCA has been applied to the constituents of the S&P500 using 30 days of daily changes in stock price as well as 6 months of weekly stock price changes to predict the most recent day’s price change (up or down). This scatter plot shows that it is not as simple as just using previous prices to predict the next day’s direction.

**References**  
[1] Aroussi, R. (n.d.). yfinance: Yahoo! Finance market data downloader. GitHub. Retrieved June 2025, from <https://github.com/ranaroussi/yfinance>  
[2] Scikit-learn documentation: <https://scikit-learn.org>  
[3] Keras documentation: <https://keras.io>