Capstone Project Proposal: Investment Trading with Yahoo Finance Data

1. **Domain Background**

Investment trading is a critical domain in finance, involving decision-making based on economic indicators, market sentiment, and statistical signals. Traditionally dominated by manual and heuristic methods, modern trading increasingly relies on machine learning (ML) models for predictive insights and automated decision-making. With the availability of historical financial data from platforms such as Yahoo Finance, there is a unique opportunity to apply ML algorithms to forecast stock price movements and improve trading strategies.

2. **Problem Statement**

The financial market is inherently volatile and non-linear, making it difficult to predict short-term price movements. The core problem to be addressed in this project is: Can we predict the next-day direction (up or down) of a stock's closing price using technical indicators derived from historical data? This is a binary classification task where the performance of the model can significantly impact trading outcomes and profitability.

3. **Solution Statement**

This project proposes the development of a machine learning classification model to predict next-day stock price direction. The solution will leverage engineered features from historical OHLCV (Open, High, Low, Close, Volume) data, including technical indicators such as MACD, RSI, SMA, and Bollinger Bands. Models such as Random Forest, XGBoost, and LSTM will be explored to capture both non-linear relationships and temporal dependencies. The final solution will be evaluated not only on classification metrics but also on its effectiveness in simulated trading (e.g., ROI, Sharpe Ratio).

4. **Datasets and Inputs**

Source: Yahoo Finance API via yfinance Python library
Data: Daily OHLCV data for selected tickers (e.g., AAPL, AMZN, SPY) over the past 5 years
Features:
Raw: Open, High, Low, Close, Volume
Engineered: MACD, RSI, SMA, EMA, Bollinger Bands, Daily Returns
Target: Binary indicator if the next-day closing price is higher than today
Preprocessing: Data cleaning, feature normalization, handling of missing values, and temporal train-test split

5. **Benchmark Model**

The benchmark model will be a Naïve Classifier that always predicts the direction of the previous day's movement (i.e., if price went up today, it will predict it goes up tomorrow). This model will help establish a baseline accuracy (~50-55%) against which more sophisticated models can be compared.

6. Evaluation Metrics

Classification Metrics:
Accuracy,
Precision, Recall, F1-Score
Confusion Matrix

Trading Performance Metrics:
ROI (Return on Investment)
Sharpe Ratio (Risk-adjusted return)
Max Drawdown

7. Project Design

- Data Collection: Use yfinance to extract data for selected stocks
- Feature Engineering: Compute technical indicators and create binary target
- EDA: Visualize price trends, correlation matrices, and indicator behavior
- Modeling:
    - Baseline: Logistic Regression / Random Forest
    - Advanced: XGBoost, LSTM
- Validation: k-fold cross-validation and test set evaluation
- Backtesting: Simulate trades using model predictions
-
- Deployment:
    - Option 1: Deploy with interactive predictions
    - Option 2: Publish a detailed technical blog post documenting findings