**S&P 500 주가 데이터 분석 및 예측 시스템 보고서**

**1. 목적 및 배경**

주가 패턴 분석 및 예측을 위하여 S&P 500 데이터의 6개월 데이터를 SQ Lite 데이터베이스에 저장 후 정제

**2. 데이터 개요**

**2.1 데이터 소스**

* **Yahoo Finance API (yfinance)**
  + S&P 500에 포함된 종목들의 실시간 및 과거 주가 데이터를 수집하는 데 사용됨.
  + FinanceDataReader를 사용하여 S&P 500 종목 리스트를 확보.

**2.2 주요 컬럼**

주가 데이터는 다음과 같이 구성된다.

|  |  |
| --- | --- |
| **Column** | **설명** |
| ticker | 주식 종목의 티커 심볼 (예: AAPL, MSFT) |
| date | 주가 기록 날짜 |
| Open | 당일 시가 |
| High | 당일 최고가 |
| Low | 당일 최저가 |
| Close | 당일 종가 |
| Volume | 거래량 |

**2.3 대상 데이터**

* S&P 500 지수에 포함된 모든 종목
* 최근 6개월 간의 데이터를 기준으로 학습 진행
* 10 거래일(2주일) 예측 수행

**3. 데이터 전처리 목표**

주가 데이터는 금융 데이터의 특성을 고려하여 아래와 같은 전처리 과정을 거친다.

**3.1 데이터 변환**

* Close(종가) 값을 표준화(Standardization)하여 모델 학습에 사용
* 주식의 티커(Ticker Symbol)를 카테고리별 로One-Hot Encoding 처리하여 학습에 반영

**3.2 중복 제거**

* SQLite 데이터베이스에 기존 데이터가 존재하는 경우 중복 제거 후 갱신
* 최신 6개월 데이터만 유지하여 데이터베이스 최적화 수행

**3.3 데이터 저장**

* SQLite 데이터베이스 (predictions.db)를 활용하여 데이터 저장
  + stock\_data 테이블: 원본 주가 데이터 저장
  + stock\_predictions 테이블: 예측 결과 저장

**4. 데이터 처리 단계**

**4.1 데이터 수집**

* yfinance 라이브러리를 활용하여 S&P 500 종목들의 주가 데이터를 수집
* 각 종목에 대해 최근 6개월 간의 데이터를 가져옴
* SQLite에 저장하여 효율적인 검색 및 업데이트 가능하도록 관리

**4.2 데이터 필터링**

* 종목별 최신 데이터만 유지하여 중복을 방지
* 거래량(Volume)이 0인 경우 해당 날짜 데이터는 제외

**4.3 데이터 정제**

* StandardScaler를 사용하여 주가 데이터 정규화
* OneHotEncoder를 사용하여 카테고리 별 인코딩 진행

**4.4 데이터 저장**

* SQLite 데이터베이스를 사용하여 실제 주가 데이터 및 예측값 저장
* 종목별 예측 결과를 stock\_predictions 테이블에 저장
* 중복 데이터 갱신 (INSERT OR REPLACE 쿼리 적용)

**5. 결과 및 요약**

**1. 처리 전 데이터**

* Yfinace를 통해 수집한 주가데이터

**2. 처리 후 데이터**

* 종가(Close)에 대한 값 스케일링(StandaraSclaer)
* 티커 카테고리 별 OnehotEncoding
* SQLite 데이터베이스 테이블
  + stock\_data : 6개월 주가 데이터