서지 정보

논문 제목	데이터 웨어하우징과 데이터 마이닝을 통한 코스피 일별 변동 및 개별 기업 주가 변동 예측 모형 개발			
참가 전형	일반인		응모 주제	산업 경제 일반
소속 기관	이름	연구참여형태	소속 기관	
	김덕현	주저자	경상대학교 일반대학원 경영정보학과	
	황진경	교신저자	효정산업 연구원	
	조승현	공저자	경상대학교 자연과학대학 컴퓨터과학과	
연락처	이름	H.P	E-mail	
	김덕현	010-6627-1654	kdh294@gnu.ac.kr	
	황진경	010-4752-7707	davidhwangjg@gmail.com	
	조승현	010-7557-4898	gmelanals@gmail.com	

데이터 웨어하우징과 데이터 마이닝을 통한 코스피 일별 변동 및 개별 기업 주가 변동 예측모형 개발

김덕현, 조승현, 황진경

<목 차>

Ⅰ. 서론
Ⅱ. 이론적 배경
Ⅲ. 실험 설계
Ⅳ. 연구 결과
Ⅴ. 결론
참고문헌

Ⅰ. 서론

현재 지구촌은 유례없는 팬데믹을 야기한 코로나 바이러스 감염증-19(COVID-19, 이하 코로나19)로 인해 고통 받고 있다. 코로나19로 인한 문제는 개인뿐만 아니라 공동체, 사회, 국가 전반에 걸쳐 발생하고 있다. 한국도 마찬가지로 1, 2차 등으로 이어지는 대유행이 여전히 끝나지 않고 있으나 언젠가는 종식될 것이라고 믿어 의심치 않는다.

특히 코로나19는 국내외 산업과 증시에 큰 타격을 주었는데, 한국의 경우 코로나19의 창궐로 인한 대두된 위기의식으로 인하여 KOSPI 일별 지수가 약 10일 만에 505.29pt 하락하였다.[1] 큰 낙폭의 KOSPI 지수 하락은 IMF 외환위기 때와 비견되며, 코로나19의 여파가 얼마나 강력했는지 알 수 있다. 그러나 한국의 증시는 다행히 V자 반등에 성공하였으며, 이러한 현상은 코로나19와 직·간접적인 관련이 있는 부분과 정부 정책 발표 및 세계 경제 흐름과 관련이 있는 부분으로 나누어 현황을 파악할 수 있다.

먼저 코로나19 백신 개발에 관련된 '바이오'와 '제약', 코로나19로 인한 생활 패턴의 변화와 관련된 '언택트', '온라인 교육', '포장' 등이 수혜주로 떠올랐다. 그리고 급변하는 사회와 기술력 증진에 맞물려 4차 산업혁명의 대표적인 산업인 '5G', '전기차', '수소차', '신재생에너지', '제약 및 바이오' 등이 수혜주로 떠오르며 증시의 고공행진을 견인하였다.

그러나 앞서 언급한 수혜주의 경우, 기업의 본원적 가치에 의거한 투자가 아닌 코로나

1) 2020년 03월 10일 1,962.27pt에서 동년 동월 19일 1,457.64pt로, ▽505.29pt

19라는 큰 충격에 대한 반대급부적 기대 심리가 반영된 투자이라 판단된다. 그렇기 때문에 실물 경제 상황이 좋지 않은 작금의 상황에서, 과거 닷컴 버블로 인한 증시 폭락과 유사한 공포감이 시장 전반에 깔려 있는 실정이다.

기대와 공포가 공존하는 현 시점에서, 각종 거시 경제 지표와 국내외 주가 지표, 그리고 기타 외부 환경 요인들이 국내 증시와 어떠한 영향 관계가 있는지 살펴보는 것은 매우 중요하다고 판단된다. 이를 위해 본 연구는 코로나19로 인한 증시 변동과 그에 따른 개별 기업의 수익 여부가 어떠했는지를 파악하고자 하였으며 그 과정은 다음과 같다.

첫째, 각종 웹 사이트로부터 연구 문제 해결에 적합한 데이터를 수집하였으며, 수집된 자료를 통합하는 데이터 웨어하우징을 시도하였다. 둘째, 구축된 데이터 웨어하우스를 토대로 코스피 변동 여부와 개별 기업에 대한 주가 변동 여부에 대한 패턴을 도출하기 위해 데이터 마이닝 기법으로 분석하였다. 셋째, 데이터 마이닝 기법 중 분류분석을 시행했다. 넷째, 분석과 평가를 통해 적합한 예측모형을 선정하였다. 다섯째, 선정된 예측모형에 대한 해석과 규칙 도출, 투자 전략 제안, 논의 등을 다루었다.

본 연구의 구성은 다음과 같다. 2장에서 한국 주식시장에서의 주가 수익률과 관련된 선행 연구들을 살펴보고, 본 연구에서 활용한 변수들에 대한 이론적 근거를 파악하였다. 3장에서는 본 연구의 방법론과 분석 자료에 대한 일반적 특성, 실험에 투입한 변수 등을 정리하였다. 4장에서 실험 과정을 통해 도출된 연구 결과를 종합하고, 합리적인 투자를 위한 의사결정 규칙을 도출하였다. 5장에서는 본 연구의 결론을 제시하였다.

Ⅱ. 이론적 배경

김유신 등(2012)의 연구에서는 뉴스와 주가에 대한 영향 관계를 분석하였으며, 빅데이터 감성분석을 통한 지능형 투사의사결정모형을 제안하였다. 비정형 텍스트인 뉴스 컨텐츠를 오피니언 마이닝을 통해 분석하여, 주가지수의 등락을 예측하는 지능형 투자의사결정 모형을 제시하였다. 그 결과로 뉴스 컨텐츠의 감성분석 결과값과 주가지수의 등락은 유의한 관계를 가지고 있으며, 특히 주식시장 개장 전 뉴스들과 주가지수의 등락과의 관계가 통계적으로 유의함을 규명하였다.

김주일(2013)의 연구에서는 코스피 지수 및 코스닥 지수와 환율과의 상호연관성에 관한 분석을 진행하였다. 주요 변수로 코스피 지수, 코스닥 지수, 원/달러 환율을 활용하였으며, 인과관계분석과 충격반응분석 및 분산분해를 실시하였다. 그 결과 외국인 투자자가 국내 주식을 매매하는 비중에 따라 국내주가지수와 환율 간의 영향이 변동됨을 규명하였다.

김용재와 이상춘(2017)의 연구는 한국의 거시경제 지표가 기업의 주가수익률에 어떤 영향을 미치는가에 대한 SVAR을 활용하여 분석하였는데, SVAR이란 구조적 벡터자기회귀모형을 의미한다. 거시경제변수로 환차익(손)율과 원유가, 총통화, 3년만기 국고채수익률, 국제수지 등을 활용하여 개별 기업의 주가 수익률을 분석하였다. 분석결과로 거시경제변수 중 주가 수익률에 가장 큰 영향이 나타난 것으로 국제수지가 있으며, 이는 우리나라가 수출주도형 산업으로 외환의 주공급처가 기업임을 확인할 수 있음을 규명하였다.

박정미와 박정태(2019)의 연구에서는 ICT 관련 상호변경이 주가에 미치는 영향을 분석하였으며, 2000년의 IT 버블의 여부에 대한 파악을 하고자 하였다. 주요 변수로 주가수익률과 상호, 상호 변경여부, 해당 기업의 정량적·정성적 변수들을 활용하였다. 분석결과로 ICT 관련 상호변경은 기업 가치에 긍정적인 영향을 주고 있고, 상호변경이라는 이슈에 대해 시장이 반응하고 있음을 규명하였다. 또한 상호변경 공시효과에 대한 반응을 구체적으로 입증하였고, 상호변경 의사결정자와 상호변경 기업에 투자하려는 의사결정자에 대한 투자 지침을 제공하였다.

고강석(2018)의 연구에서는 업종별 주가지수와 주가, 환율 관계의 안정성을 규명하고자 분석을 시도하였다. 주요 변수로 업종별 주가지수와 환율 등을 활용하였으며 전통적 접근과 포트폴리오 접근의 두 가지 견해를 통해 결과를 해석하였다. 주가와 환율의 동태적 관계는 과거에 비하여 현재에는 미비한 것으로 분석 결과를 제시하였다.

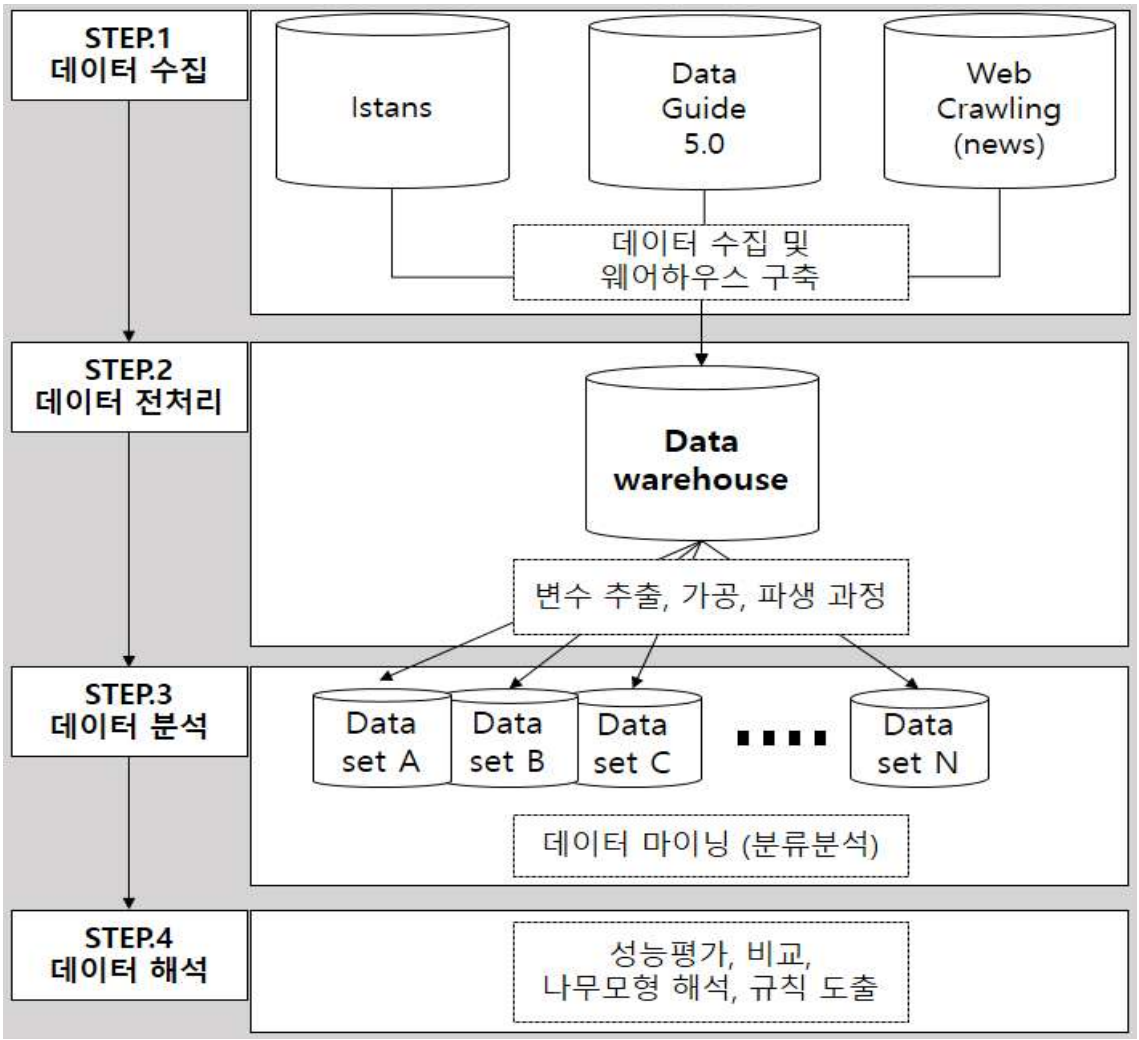
상위의 선행 연구들을 종합하면 다음과 같다. 금리, 국고채, 환율, 국제수지, 상호명, 뉴스 키워드 등의 다양한 변수들을 활용함으로 주가 수익 여부 혹은 증시에 대한 변동 여부 등을 파악하고자 하였다.

본 연구는 선행연구에서 다룬 변수들 중에서 자료의 원천이 되는 각각의 웹 데이터베이스로부터 데이터를 수집하고, 변수를 구성하여 증시 변동 여부와 개별 기업 주가 변동 여부를 파악하고자 한다.

Ⅲ. 실험 설계

3.1 연구 진행과정

본 연구의 진행과정은 [그림 1]과 같다.



[그림 1] 연구 진행과정

연구 진행과정은 크게 4 단계로 구성된다. 첫 번째, 데이터 수집 단계에서 각각의 데이터베이스나 웹으로부터 연구에 적합한 데이터를 수집하였다. 출처가 다른 데이터이기 때문에, 데이터를 축적하고 관리하는 방식이 각기 다르다. 그러므로 공통의 형식으로 변환하여 하나의 데이터웨어하우스를 구축하였고, 이러한 작업을 데이터웨어하우징이라 한다.

두 번째, 데이터 전처리 단계에서 데이터 마이닝 분석에 용이한 형태로 데이터웨어하우스 내에 포함된 데이터를 처리하였다. 공통의 양식으로 변환하여 데이터웨어하우스를 구축하였다 하더라도, 데이터와 변수의 성격에 맞게 데이터를 정제하고, 통합하고, 변환하고, 축소하는 과정이 필요하다. 본 연구에서 다루는 데이터에는 지수값, 주가, 금리, 환율, 일자, 유가 등이 주로 포함되며 수치형 데이터에 속한다. 이를 데이터의 의미와 성격에 맞게 범주화를 하거나, 전일비 형태로 데이터를 가공하는 등의 과정을 진행하였다. 결과적으로 데이터 전처리 과정을 통해 독립변수군과 목표변수군을 구성하였다.

세 번째, 데이터 분석 단계로 본 연구에서는 데이터 마이닝 분석 기법을 사용하였다. 앞선 데이터 전처리 단계를 통해 구성된 독립변수군과 목표변수군의 조합으로 이루어진 데이터 셋을 구성하였고, 각각의 데이터 셋에 대한 데이터 마이닝 분석을 시행하였다. 데이터 마이닝 분석 기법 중에서도 분류분석(classification)을 시도하였고, 분류분석의 한 분야인 의사결정나무 분석을 사용하였다. 의사결정나무(decision tree)는 독립변수와 목표변수의 관계가 '화이트 박스' 형태로 나타나기 때문에 의사결정 규칙을 발견하거나, 전략적 제안점을 도출하기 용이하다. 본 연구에서는 의사결정나무의 알고리즘으로 C4.5를 활용하였다.

네 번째, 데이터 해석 단계로 각각의 데이터 셋에 대한 의사결정나무 분석을 시행한 결과를 해석하고, 성능을 평가·비교하며, 이를 토대로 규칙을 도출하는 과정을 수행하였다. 예측모형 혹은 분류기의 성능을 평가하는 지표로 예측률, 정분류율이 있다. 평가지표를 통해 예측모형의 성능을 비교하고, 나무모형 해석과 규칙 도출에 적합한 예측모형을 선정하였다. 선정된 예측모형 중에서도 유의미한 주요 규칙을 선별하였고, 이를 바탕으로 시사점과 전략적 제안을 제시하였다.

3.2 분석 자료

Istans에서 제공하는 종합주가지수, 기업경기실사지수, 실질 경제성장률 등의 국내외 거시 경제 지표와 관련된 데이터를 수집하였다. Data Guide 5.0에서 제공하는 개별 기업의 주식 거래량, 수익률, 투자자별 순매수 현황 등을 수집하였다. 끝으로 네이버 증권에서 제공하는 메인 뉴스에 대한 텍스트 마이닝을 시도하였다. 일자별 기사의 제목과 헤드라인에 특정 키워드가 노출되면, 키워드 별로 계수하여 변수를 구성하였다.

Istans로부터 수집한 데이터는 국가별 실질 경제성장률(미국, 중국, 한국), 국가별 주가지수(미국, 한국), 한국 업종별 주가 지수, 동행종합지수, 기업 경기실사지수, 생산자·소비자 물가지수, 국고채 3년(한국), 국제금리 10년(미국), 월별 원/달러 환율, 월별 유가(두바이유, 브렌트유, 서부텍사스중질유)로 거시 경제 지표이기 때문에 데이터의 시간적

기준이 월 단위이거나 분기 단위이다.

Dataguide 5.0으로부터 수집한 데이터는 개별 기업의 일자별 종가를 수집하였으며, 삼성전자, LG화학, 현대차, 셀트리온, 네이버, 카카오, 엔씨소프트, 롯데쇼핑, 하나투어, 한국항공우주(KAI), 대한항공, 아이에스동서, 대상, 에스엠, 미래에셋대우로 구성된다. 연구진의 판단 하에 주식시장 내에서 업종을 선도하는 기업이자 코로나19로 인한 산업 전반의 영향력을 파악하기 위해 해당 기업의 일별 종가만을 활용하였다.

네이버 증권 주요 뉴스로부터 수집한 데이터는 일자별로 주요 뉴스에 게시된 기사의 제목과 헤드라인이며, '유가', '금리', '환율', '기관', '외국인', '개인', '반등', '급락', '코로나 및 우한폐렴'이라는 키워드가 포함되면 계수하였다. 코로나19의 경우 초기 발생지역이 중국 우한 지역이기 때문에, 바이러스에 대한 정식 명칭이 정해지지 않았을 때부터 일정시점까지 우한폐렴과 코로나19가 혼재되어 사용되었다. 그러므로 이를 동의어로 간주하고 함께 계수하였다. 텍스트 마이닝의 도구로 python 3.8.5 버전을 활용했고, KoNLPy(한국어 자연어처리 라이브러리)를 활용하였다(박은정, 조성준, 2014).

코로나19로 인한 증시 변동과 그에 따른 개별 기업의 수익 여부를 파악하기 위해, 본 연구진은 2018년 01월 01일부터 2020년 06월 30일까지의 기간 내에 포함되는 데이터를 수집하였다. 과거의 데이터 수집에 대한 어려움은 없었으나, 비교적 최근인 2020년 2분기 혹은 2020년 04월 01일부터 06월 30일에 해당하는 분기별, 월별 데이터에 대한 데이터 수집의 어려움이 존재했다. 이를 보완하기 위해 한국은행, 한국경제연구원 등의 공시 자료를 활용하여 거시 경제 지표와 주식시장 개장 여부, 일별 코스피, 코스닥 지수 등을 누락된 정보를 추가하였다.

상위 연구 진행과정에서 언급한 바와 같이, 대부분의 수집 자료의 유형이 수치형이기 때문에 전일(前日)·전월(前月)·전분기(前分期) 대비 상승 및 하락 여부를 판단하여 범주화하거나, 산술평균을 기준으로 범주화하는 데이터 전처리 과정을 거쳤다. 그 이유는 시계열적 성격을 포함하고, 수치적 상승 혹은 하락에 따른 증시 변동이나 수익 여부를 파악하고, 이를 통해 합리적 투자 의사결정을 지원하기 위해 수치형 데이터를 범주형 데이터로 대부분 가공하였다.

3.3 변수

상위의 과정을 통해 구축된 데이터웨어하우스를 바탕으로 분석을 위해 추출한 변수는 총 100개이며 실험1에서 목표변수 포함 99개의 변수를 활용하였고, 실험2에서 목표변수 포함 86개의 변수를 활용하였다. 즉, 실험1과 2를 진행할 때 목표변수에 따라 독립변수 구성을 달리 하였다.

실험1의 목표변수는 KOSPI 일별 변동이며, KOSPI 월별 변동을 제외하고 추출한 모든 변수를 독립변수로 활용하였다. 실험2의 목표변수는 개별 기업의 일별 종가 변동이다. 예를 들어, 삼성전자 일별 종가 변동을 목표변수로 설정하였을 때, 타기업의 일별 종가 변동 변수는 독립변수로 활용하지 않았다.

본 연구에서 활용한 변수에 대한 정리는 <표 1>과 같이 요약하였다.

<표 1>. 독립변수와 목표변수 요약

코드명	변수 설명	자료 원천	시점	변환 여부	변환 기준	실험1	실험2	참고
day	요일	X	일	O	해당일자	O	O	월~금
week	주	X	주	O	해당일자	O	O	해당년도 해당월 해당주
quarter	분기	X	분기	O	해당일자	O	O	해당년도 해당분기
month	월	X	월	O	해당일자	O	O	1월~12월
USA_egr	미국 경제성장률	Istans	분기	O	전분기비 초과/미만	O	O	기간 중 실질경제성장률
CHN_egr	중국 경제성장률	Istans	분기	O	전분기비 초과/미만	O	O	기간 중 실질경제성장률
KOR_egr	한국 경제성장률	Istans	분기	O	전분기비 초과/미만	O	O	기간 중 실질경제성장률
KOR_long_ir	한국 국고채 (3년)	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균금리
USA_long_ir	미국 장기국채 (10년)	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균금리
KOR_short_ir	한국은행 기준금리	BOK	일	X	-	O	O	기간 중 평균금리
interest rate	금리 변경 여부	BOK	일	O	전월비 초과/미만	O	O	금리 변경일 표기
exchange_ monthly	환율(원/달러) 월별 변동	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균환율
exchange_ daily	환율(원/달러) 월별 변동	data guide	일	O	전일비 초과/미만	O	O	일별 환율
oilprice	원유 평균가 월별 변동	Istans	월	O	전월비 초과/미만	O	O	월별 원유 평균가
DowJones_ quarterly	Dow Jones 분기별 변동	Istans	분기	O	전분기비 초과/미만	O	O	다우존스 분기별 평균지수
NASDAQ_ quarterly	NASDAQ 분기별 변동	Istans	분기	O	전분기비 초과/미만	O	O	나스닥 분기별 평균지수
DowJones65_ daily	Dow Jones 65 composite 일별 변동	data guide	일	O	전일비 초과/미만	O	O	다우존스 65 일별 지수
NASDAQ_ daily	NASDAQ 일별 변동	data guide	일	O	전일비 초과/미만	O	O	나스닥 일별 지수
KOSDAQ_ daily	KOSDAQ 일별 변동	data guide	일	O	전일비 초과/미만	O	O	코스닥 일별 지수
STI_1	제조업 주가지수	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균주가지수
STI_2	음식료품 주가지수	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균주가지수

STI_3	섬유의복 주가지수	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균주가지수
STI_4	종이목재 주가지수	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균주가지수
STI_5	화학 주가지수	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균주가지수
STI_6	의약품 주가지수	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균주가지수
STI_7	비금속광물 주가지수	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균주가지수
STI_8	철강금속 주가지수	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균주가지수
STI_9	기계 주가지수	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균주가지수
STI_10	전기전자 주가지수	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균주가지수
STI_11	의료정밀 주가지수	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균주가지수
STI_12	운수장비 주가지수	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균주가지수
STI_13	유통업 주가지수	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균주가지수
STI_14	전기가스업 주가지수	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균주가지수
STI_15	건설업 주가지수	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균주가지수
STI_16	운수창고업 주가지수	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균주가지수
STI_17	통신업 주가지수	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균주가지수
STI_18	금융업 주가지수	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균주가지수
STI_19	은행 주가지수	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균주가지수
STI_20	증권 주가지수	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균주가지수
STI_21	보험 주가지수	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균주가지수
STI_22	서비스업 주가지수	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균주가지수
PPI	생산자 물가 총지수	Istans	월	O	전월비 초과/미만	O	O	2015년＝100
CPI	소비자 물가 총지수	Istans	월	O	전월비 초과/미만	O	O	2015년＝100
CCI	동행종합지수	Istans	월	O	전월비 초과/미만	O	O	2015년＝100
A_BSI1	전산업 업황실적BSI	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균업황실적
A_BSI2	전산업 매출실적BSI	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균매출실적
A_BSI3	전산업 자금사정실적BSI	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균자금사정실적
B_BSI1	제조업 업황실적BSI	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균업황실적
B_BSI2	제조업	Istans	월	O	전월비	O	O	기간 중

	매출실적BSI				초과/미만			평균매출실적
B_BSI3	제조업 자금사정실적BSI	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균자금사정실적
C_BSI1	대기업 업황실적BSI	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균업황실적
C_BSI2	대기업 매출실적BSI	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균매출실적
C_BSI3	대기업 자금사정실적BSI	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균자금사정실적
D_BSI1	중소기업 업황실적BSI	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균업황실적
D_BIS2	중소기업 매출실적BSI	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균매출실적
D_BSI3	중소기업 자금사정실적BSI	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균자금사정실적
E_BSI1	중화학공업 업황실적BSI	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균업황실적
E_BSI2	중화학공업 매출실적BSI	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균매출실적
E_BSI3	중화학공업 자금사정실적BSI	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균자금사정실적
F_BSI1	경공업 업황실적BSI	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균업황실적
F_BSI2	경공업 매출실적BSI	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균매출실적
F_BSI3	경공업 자금사정실적BSI	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균자금사정실적
G_BSI1	수출기업 업황실적BSI	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균업황실적
G_BSI2	수출기업 매출실적BSI	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균매출실적
G_BSI3	수출기업 자금사정실적BSI	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균자금사정실적
H_BSI1	내수기업 업황실적BSI	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균업황실적
H_BIS2	내수기업 매출실적BSI	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균매출실적
H_BSI3	내수기업 자금사정실적BSI	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균자금사정실적
I_BSI1	비제조업 업황실적BSI	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균업황실적
I_BSI2	비제조업 매출실적BSI	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균매출실적
I_BSI3	비제조업 자금사정실적BSI	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균자금사정실적
J_BSI1	서비스업 업황실적BSI	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균업황실적
J_BSI2	서비스업 매출실적BSI	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균매출실적
J_BSI3	서비스업 자금사정실적BSI	Istans	월	O	전월비 초과/미만	O	O	기간 중 평균자금사정실적
word_oil	유가 키워드	naver news	일	O	산술평균 초과/미만	O	O	기간 중 뉴스 상 노출빈도
word_interest	금리 키워드	naver news	일	O	산술평균 초과/미만	O	O	기간 중 뉴스 상 노출빈도

word_exchange	환율 키워드	naver news	일	O	산술평균 초과/미만	O	O	기간 중 뉴스 상 노출빈도
word_retail	개인 키워드	naver news	일	O	산술평균 초과/미만	O	O	기간 중 뉴스 상 노출빈도
word_institution	기관 키워드	naver news	일	O	산술평균 초과/미만	O	O	기간 중 뉴스 상 노출빈도
word_foreign	외국인 키워드	naver news	일	O	산술평균 초과/미만	O	O	기간 중 뉴스 상 노출빈도
word_dcbounce	반등 키워드	naver news	일	O	산술평균 초과/미만	O	O	기간 중 뉴스 상 노출빈도
word_crashed	급락 키워드	naver news	일	O	산술평균 초과/미만	O	O	기간 중 뉴스 상 노출빈도
word_corona	코로나 키워드	naver news	일	O	산술평균 초과/미만	O	O	기간 중 뉴스 상 노출빈도
Samsung	삼성전자 일별 종가 변동	data guide	일	O	전일비 초과/미만	O	T.V1	반도체 및 전자
LG	LG화학 일별 종가 변동	data guide	일	O	전일비 초과/미만	O	T.V2	화학
Hyundai	현대차 일별 종가 변동	data guide	일	O	전일비 초과/미만	O	T.V3	자동차
Celltrion	셀트리온 일별 종가 변동	data guide	일	O	전일비 초과/미만	O	T.V4	바이오
Naver	NAVER 일별 종가 변동	data guide	일	O	전일비 초과/미만	O	T.V5	IT
Kakao	카카오 일별 종가 변동	data guide	일	O	전일비 초과/미만	O	T.V6	IT
Ncsoft	엔씨소프트 일별 종가 변동	data guide	일	O	전일비 초과/미만	O	T.V7	게임
Lotteshop	롯데쇼핑 일별 종가 변동	data guide	일	O	전일비 초과/미만	O	T.V8	유통
Hanatour	하나투어 일별 종가 변동	data guide	일	O	전일비 초과/미만	O	T.V9	관광
KAI	한국항공우주 일별 종가 변동	data guide	일	O	전일비 초과/미만	O	T.V10	방산
Koreanair	대한한공 일별 종가 변동	data guide	일	O	전일비 초과/미만	O	T.V11	항공
isDongseo	아이에스동서 일별 종가 변동	data guide	일	O	전일비 초과/미만	O	T.V12	건설업 및 폐기물
Daesang	대상 일별 종가 변동	data guide	일	O	전일비 초과/미만	O	T.V13	식료품
SM	에스엠 일별 종가 변동	data guide	일	O	전일비 초과/미만	O	T.V14	엔터테인먼트
maDaewoo	미래에셋대우 일별 종가 변동	data guide	일	O	전일비 초과/미만	O	T.V15	금융
KOSPI_monthly	KOSPI 월별 변동	Istans	월	O	전월비 초과/미만	X	O	코스피 월별 평균지수
KOSPI_daily	KOSPI 일별 변동	data guide	일	O	전일비 초과/미만	T.V	O	코스피 일별 지수

<표 1>은 본 연구에서 활용한 변수의 코드명과 해당 변수에 대한 설명, 자료의 원천이 차례로 명시되어 있다. 해당 변수가 일별 데이터를 포함하는지, 월별 데이터를 포함하는지 등을 시점이라는 항목을 통해 명시하였다. 데이터 전처리 과정에서 수치형 데이터를 범주형 데이터로 바꾸었는지 여부는 변환여부를 통해 명시하였고, 변환의 기준도 함께 정리하였다.

실험1과 실험2에서 T.V란 목표변수(target variable)을 의미하며, 실험1의 경우 목표변수 1개, 독립변수 98개를 투입・분석하여 KOSPI 일별 변동 예측모형을 구축하였다. 실험2의 경우 개별 기업의 일별 종가 변동이 목표변수로 활용되기 때문에 목표변수 1개와 독립변수 85개를 투입・분석하여 각각의 일별 종가 변동 예측모형을 구축하였다.

Ⅳ. 연구 결과

4.1 연구 자료의 일반적 특성

본 연구는 2018년 01월 01일부터 2020년 06월 30일까지의 기간을 중심으로 각종 거시경제지표와 증시, 개별 주식의 종가를 수집하였다. 해당 기간은 총 913일로, 휴일 등의 이유로 주식 시장이 개장하지 않은 300일을 제외하였다. 즉, 613개의 인스턴스(instance)를 확정하였고, 실험에 따라 투입된 변수(feature)가 각기 다르다.

최종적으로 선정된 인스턴스의 일반적인 특성은 <표 2>와 같다. <표 2>에 주요 변수의 하위집단 분포와 해당 기간의 일별 주가지수에 대한 통계량 등을 정리하였다.

먼저 목표변수 군으로 구분한 변수들의 하위집단 분포를 살펴보면, 해당 기간동안 전일대비 종가 가격의 상승 횟수가 그렇지 않은 경우보다 많은 기업에 삼성전자, LG화학, 엔씨소프트, 에스엠이 속한다. 나머지 기업들은 전일대비 종가 가격이 동일하거나 하락한 횟수가 많은 경우에 속한다. 코스피 일별 변동의 경우 전일대비 코스피 지수가 상승한 경우가 331회로, 동일하거나 하락한 경우인 282회보다 많았다. 특히 코로나19로 인한 항공업과 관광업에 대한 타격이 컸음을 상기할 때, 현실이 반영된 결과라 판단할 수 있다.

다음으로 키워드 변수들은 유가, 금리, 환율, 개인, 기관, 외국인, 반등, 급락, 코로나라는 단어가 네이버 증권 메인 뉴스에 나타난 빈도를 일자별로 계수・합산하여 데이터를 구성하였다. 모든 변수의 최솟값은 0이며, 메인 뉴스에 해당 단어가 노출되지 않았음을 의미한다. 해당기간 동안 금리, 환율, 기관, 외국인이라는 단어가 포함된 기사가 메인 뉴스에 비교적 많이 노출되었던 것으로 보인다.

<표 2>. 연구 자료의 일반적 특성

목표변수군			독립변수군					
코드명	전일비 초과(회)	전일비 이하(회)	코드명	평균 (회)	하위집단		최대	최소
					Y	N		
Samsung	307	306	word_oil	1.40	210	403	16	0
LG	310	303	word_interest	2.73	240	373	27	0
Hyundai	258	355	word_exchange	2.17	359	254	12	0
Celltrion	294	319	word_retail	1.05	362	251	11	0
Naver	285	328	word_institution	2.08	344	269	10	0
Kakao	288	325	word_foreign	2.83	305	308	13	0
NCsoft	305	308	word_dcbounce	2.27	289	324	15	0
Lotteshop	255	358	word_crashed	1.80	234	379	21	0
Hanatour	281	332	word_corona	1.82	107	506	36	0
KAI	287	326	해당기간 국내·외 일별 주가지수 (단위: pt)					
Koreanair	277	336						
isDongseo	299	314	코드명	산술평균	표준편차		최대	최소
Daesang	272	341	DowJones65_daily	8510.28	±538.35		9710.01	6100.31
SM	302	311	NASDAQ_daily	7933.61	±768.60		10131.37	6192.92
maDaewoo	286	327	KOSDAQ_daily	726.67	±92.99		927.05	428.35
KOSPI_daily	331	282	KOSPI_daily	2175.21	±187.32		2598.19	1457.64

특히 코로나라는 단어가 포함된 기사는 일일 최대 36회까지 메인 뉴스에 노출되었다. 평균인 1.82회보다 적게 노출된 경우를 'N'으로 많게 노출된 경우를 'Y'로 하위집단을 구분하였으며, N집단에 속하는 인스턴스가 506개로 월등히 많음에도 불구하고 짧은 기간동안 많은 수의 기사가 작성되고 메인 뉴스에 노출되었음을 유추해볼 수 있다.

끝으로 국내·외 일별 주가지수의 통계량을 간단히 정리하였다. 전반적으로 미국의 증시 변동폭이 한국의 증시 변동폭보다 큼을 알 수 있고, 국내로 한정할 때 코스피 지수 변동이 코스닥 지수 변동보다 큼을 알 수 있다.

4.2 예측모형 성능 평가 및 비교

실험1과 실험2를 진행하여 구축된 예측모형은 총 16개로 예측모형에 대한 성능 평가 및 비교를 <표 3>을 통해 정리하였다. 분석의 도구로 Weka 3.8.3을 활용하였고, 의사결정나무 알고리즘 중 하나인 C4.5를 활용하였다. 학습 데이터와 검정 데이터의 분할 비율로 7:3으로 설정하여 분석을 시행하였다.

먼저 예측률을 중심으로 살펴보면, 실험1과 실험2 모두를 통틀어 '코스피 일별 변동 예측모형'이 가장 높은 예측률인 78.8043%를 기록하였다. 실험2에 해당하는 개별 기업 주가 변동 예측모형 중에서 가장 높은 예측률을 기록한 모형은 '삼성전자 주가 변동

예측모형’으로 77.1739%의 예측률을 기록하였다. 가장 낮은 예측률을 기록한 모형은 ‘엔씨소프트 주가 변동 예측모형’으로 51.6304%의 예측률을 기록하였다.

<표 3>. 예측모형의 성능 평가 및 비교

실험 구분	목표변수 (코드명)	독립변수 (개)	예측률 (%)	정분류율			요약	선정
				초과	이하	가중평균		
실험1	KOSPI_daily	99	78.8043	0.765	0.814	0.788	예측률 전체 1위	✓
실험2	Samsung	85	77.1739	0.761	0.781	0.772	실험2 중 예측률 상위 1위	✓
실험2	LG	85	70.1087	0.615	0.795	0.701	실험2 중 예측률 상위 3위	
실험2	Hyundai	85	54.8913	0.354	0.706	0.549	예측률 하위 3위	
실험2	Celltrion	85	58.1522	0.616	0.551	0.582	코로나 키워드 분지기준 포함	✓
실험2	Naver	85	61.4130	0.495	0.731	0.614	-	
실험2	Kakao	85	66.8478	0.578	0.755	0.668	실험2 중 예측률 상위 4위	
실험2	NCsoft	85	51.6304	0.647	0.404	0.516	예측률 하위 1위	
실험2	Lotteshop	85	57.0652	0.462	0.630	0.571	-	
실험2	Hanatour	85	60.8696	0.627	0.596	0.609	코로나 키워드 분지기준 포함	✓
실험2	KAI	85	57.0652	0.616	0.531	0.571	-	
실험2	Koreanair	85	63.0435	0.506	0.747	0.630	-	
실험2	isDongseo	85	65.2174	0.705	0.596	0.652	실험2 중 예측률 상위 5위	
실험2	Daesang	85	53.2609	0.451	0.598	0.533	예측률 하위 2위	
실험2	SM	85	61.4130	0.687	0.554	0.614	-	
실험2	maDaewoo	85	71.7391	0.756	0.686	0.717	실험2 중 예측률 상위 2위	✓

다음으로 정분류율을 중심으로 살펴보면, 목표변수의 하위집단에 대한 정분류율과 가중평균 정분류율로 구분하여 ‘해당 예측모형이 어떤 하위집단을 비교적 잘 분류하는가?’를 파악할 수 있다. 최고 예측률을 기록한 ‘코스피 일별 변동 예측모형’의 정분류율에서는 전일 대비 동일 혹은 하락(이하)에 대한 분류가 잘 이루어짐을 알 수 있으며, 0.814를 기록하였다. ‘삼성전자 일별 변동 예측모형’의 정분류율에서도 마찬가지로 전일 대비 동일 혹은 하락(이하)에 대한 분류가 잘 이루어짐을 알 수 있으며, 0.781을 기록하였다. 반면 ‘셀트리온 일별 변동 예측모형’의 정분류율과 ‘하나투어 일별 변동 예측모형’의 정분류율은 전일 대비 상승(초과)에 대한 분류가 잘 이루어짐을 알 수 있으며, 각각 0.616과 0.627을 기록하였다. 종합하자면 개별 예측모형의 목표변수 하위집단 중 ‘이하 집단’에 대한 정분류율이 ‘초과 집단’보다 높은 경우가 9개 예측모형에서 발생하였다고, 반대로 ‘초과 집단’에 대한 정분류율이 ‘이하 집단’보다 높은 경우가 7개 예측모형에서 발생하였다.

본 연구는 코로나19가 증시 변동과 개별 주가 변동에 대한 영향이 존재하는가를 규명하는 것이 연구 목표였기 때문에, 16개의 예측모형 중에서 이에 적합한 예측모형을 선

정하였다. 먼저 예측률이 가장 높았고, 증시 변동을 직접 살펴볼 수 있는 실험1의 ‘코스피 일별 변동 예측모형’을 선정하였다. 실험2의 예측모형 중에서는 ‘삼성전자’, ‘미래에셋대우’, ‘LG화학’, ‘셀트리온’, ‘하나투어’를 선정하였다.

다음으로 선정된 예측모형 6개의 예측모형에 대한 나무모형 해석과 규칙 도출을 통해 ‘어떠한 결정요인에 의해 변동 여부가 결정되는가?’, ‘코로나19가 뉴스 기사에 노출되는 영향이 있는가?’, ‘합리적인 투자를 위한 전략은 무엇인가?’에 대한 해답을 제시하고자 한다.

4.3 선정된 예측모형의 변수 중요도

선정된 예측모형 6개의 변수 중요도를 살펴보기 위해 이득 비를 통한 변수 중요도를 산정하였다. 다양한 변수들이 투입되었기 때문에, 상위 7개를 선정하여 <표 4>에 요약하였다.

<표 4>. 선정된 예측모형의 변수 중요도

구분	실험1	실험2					참고
	코스피	삼성전자	미래에셋대우	LG화학	셀트리온	하나투어	
상위 1위	삼성전자	코스피 일별 변동	코스피 일별 변동	코스피 일별 변동	코스닥 일별 변동	코스닥 일별 변동	1) 이득비(gain ratio)는 목표변수와 단일 독립변수 간의 상관관계만을 살펴보기 때문에, 서로 독립적임. 2) 코스피가 일별 자료이기 때문에 개별 주식의 등락과 큰 상관관계가 있는 것으로 파악됨 3) 이 밖에도 금리, 주, 전기전자 주가지수, 제조업 주가지수, 일별 다우존스 65, 일별 나스닥 순으로 코스피 일별 변동에 주요한 영향을 줌
상위 2위	코스닥 일별 변동	코스닥 일별 변동	코스닥 일별 변동	코스닥 일별 변동	코스피 일별 변동	코스피 일별 변동	
상위 3위	미래에셋대우	금리 변경 여부	주	금리 변경 여부	주	금리 변경 여부	
상위 4위	아이에스 동서	주	화학 주가지수	주	의약품 주가지수	주	
상위 5위	LG화학	전기전자 주가지수	급락 키워드	급락 키워드	다우존스65 일별 변동	운수창고업 주가지수	
상위 6위	대한항공	제조업 주가지수	다우존스65 일별 변동	나스닥 일별 변동	나스닥 일별 변동	다우존스65 일별 변동	
상위 7위	한국 한공우주	급락 키워드	제조업 주가지수	화학 주가지수	제조업 주가지수	유가 키워드	

<표 4>를 살펴보면, 해당 기간 코스피 시장의 변동은 삼성전자가 가장 높은 상관관계가 있는 것으로 파악된다. 다음으로 코스닥 일별 변동이 중요한 것으로 나타났다. 다양한 기업들이 변수 중요도에 나타난 이유는 개별 주식의 일별 등락이 코스피 시장의 변동과 밀접한 연관이 있기 때문으로 파악되며, 인과적인 관계가 아니므로 본 연구에서 수집하지 않은 ‘타 기업의 일별 변동 여부’가 실험에 투입될 때 중요도의 변동 가능

성을 배제하지 않을 수 없다.

모든 개별 기업의 주가 일별 변동은 코스피 일별 변동 및 코스닥 일별 변동과 중요한 상관관계가 있는 것으로 나타났다. 셀트리온과 하나투어의 경우에는 코스닥 일별 변동이 가장 상위에 자리매김하였다.

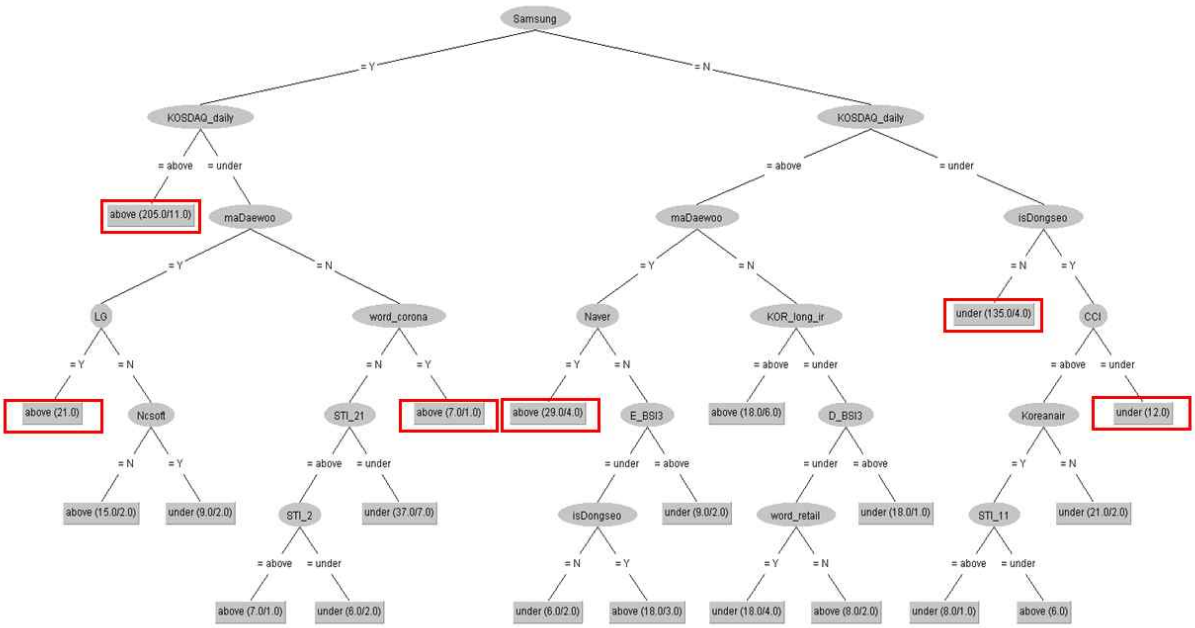
변수 중요도 순위가 높다고 해서, 반드시 나무모형 상에 분지기준으로 해당 변수가 나타나는 것은 아니다. 또한 인과적 관계가 아니기 때문에, ‘A→B’라는 독립-종속 영향관계로 볼 수 없다. 각 예측모형의 목표변수와 독립변수의 ‘독립적 상관관계’가 어떻게 나타났는지에 대한 정보를 개략적으로 얻을 수 있다.

이를 참고하여, 다음 절에서는 선정된 예측모형의 나무모형 해석과 주요 규칙 도출을 시도하고자 한다.

4.4 나무모형 해석과 주요 규칙 도출

가. 코스피 일별 변동 예측모형

코스피 일별 변동 예측모형의 분석 결과인 나무모형은 [그림 2]와 같다.



[그림 2] 코스피 일별 변동 의사결정나무

[그림 2]는 코스피 일별 변동 의사결정나무이다. 다양한 변수들이 결정요인으로 등장하였으며, 주요 규칙은 빨간색 상자로 표시하였다. 이를 토대로 도출한 코스피 일별 변

동 예측모형의 주요 규칙은 <표 5>와 같다.

1번 규칙은 삼성전자가 전일비 상승하였고, 코스닥 지수가 전일비 상승하였을 때, 코스피 지수 상승함을 의미한다. 1번 규칙은 현재 한국 주식시장의 현황을 잘 나타내는 규칙으로 판단된다. 그 이유로 삼성전자가 코스피 시장의 시가 총액 1위에 위치하고 있으며, 코스닥 지수 또한 코스피 지수의 변동과 유사한 흐름으로 변동되고 있음을 거론할 수 있다.

2번 규칙은 삼성전자가 전일비 상승하였고, 코스닥 지수가 전일비 하락하였으며, 미래에셋대우가 전일비 상승하였고, LG화학이 전일비 상승하였을 때 코스피 지수가 상승함을 의미한다.

<표 5>. 코스피 일별 변동 주요 규칙

구분	규칙 설명	규칙 (전일비)	예측수	적중률
1	Samsung = Y & KOSDAQ_daily = above	above	205	94.6%
2	Samsung = Y & KOSDAQ_daily = under & maDaewoo = Y & LG = Y	above	21	100%
3	Samsung = Y & KOSDAQ_daily = under & maDaewoo = N & word_corona = Y	above	7	85.7%
4	Samsung = N & KOSDAQ_daily = above & maDaewoo = Y & naver = Y	above	29	86.2%
5	Samsung = N & KOSDAQ_daily = under & isDongseo = N	under	135	97.0%
6	Samsung = N & KOSDAQ_daily = under & isDongseo = Y & CCI = under	under	12	100%

3번 규칙은 삼성전자가 전일비 상승하였고, 코스닥 지수가 전일비 하락하였으며, 미래에셋대우가 전일비 하락하였고, 코로나 키워드가 뉴스 기사로 노출되었을 때 코스피 지수가 상승함을 의미한다.

2번과 3번 규칙의 시사점은 다음과 같다. 코로나19로 인한 증권개설 및 주식 투자 관심이 많아졌고, 전기차에 대한 투자 관심도가 높아졌던 상황을 대변하는 규칙이라 판단된다. 자료 수집 과정에서 미래에셋대우를 금융 및 증권의 대표주로 선정하였고, LG화학을 화학 및 전기차 관련 대표주로 선정하였는데, 본 연구진의 기대에 상응하는 결과로 해석할 수 있다. 또한 코로나 키워드가 뉴스 상에 많이 노출되었을 때 미래에셋대우가 하락할 지라도 코스피 지수는 증가함을 살펴볼 수 있다. 그러므로 코로나19와 증시의 관계가 반드시 부정적이라고 볼 수 없으나 예측수가 적기 때문에 매우 중요한 규칙이라고는 볼 수 없다.

4번 규칙은 삼성전자가 전일비 하락하였고, 코스닥 지수가 전일비 상승하였으며, 미래에셋대우가 전일비 상승하였고, 네이버가 전일비 상승하였을 때 코스피 지수가 상승함을 의미한다.

4번 규칙의 시사점은 다음과 같다. 삼성전자는 하락하였으나 코스닥 지수가 상승하였을 때, 미래에셋대우를 중심으로 분지된다. 미래에셋대우가 상승하고, 네이버가 상승했을 때 코스피 지수가 상승하였다. 네이버는 쇼핑, 클라우드, 페이, 웹툰 등의 언택트 서비스 및 컨텐츠를 제공하는 기업이며, 사회적 거리두기 및 코로나19 공포로 인한 비대면 생활이 많아지면서 실적이 상승한 기업으로 거론된다.

5번 규칙은 삼성전자가 전일비 하락하였고, 코스닥 지수가 전일비 하락하였으며, 아이에스동서가 전일비 하락하였을 때 코스피 지수가 하락함을 의미한다.

6번 규칙은 삼성전자가 전일비 하락하였고, 코스닥 지수가 전일비 하락하였으며, 아이에스동서가 전일비 상승하였고, 동행종합지수가 하락하였을 때 코스피 지수가 하락함을 의미한다.

5번과 6번 규칙의 시사점은 다음과 같다. 데이터 수집 과정에서 아이에스동서는 건설업 및 폐기물 처리와 관련된 대표주로써 선정하였다. 코로나19을 진단하기 위한 진단키트에 대한 폐기물량이 급격하게 증가하였고, 폐기물 처리에 대한 관심이 높아졌다. 아이에스동서는 팬데믹 이후 동일업종 기업에 비해 수익률이 높아졌는데, 그 이유로 폐기물 처리 관련 사업을 영위함으로 시장의 관심과 기대에 부응한 것으로 보고되고 있다.

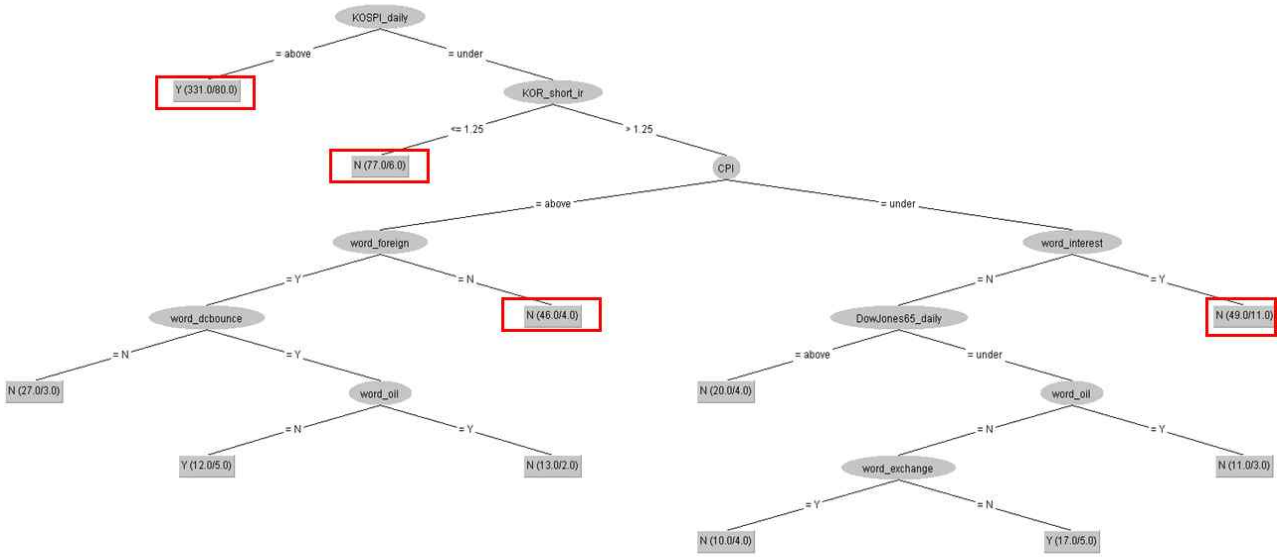
코스피 일별 예측모형의 주요 규칙들 모두 적중률이 85% 이상이다. 주요 규칙을 활용함으로 코스피 일별 변동을 파악하고 의사결정을 하는데 도움을 줄 수 있다고 판단된다.

나. 삼성전자 일별 변동 예측모형

삼성전자 일별 변동 예측모형의 분석 결과인 나무모형은 [그림 3]과 같다.

[그림 2]는 코스피 일별 변동 의사결정나무이다. 다양한 변수들이 결정 요인으로 등장하였으며, 주요 규칙은 빨간색 상자로 표시하였다. 이를 토대로 도출한 삼성전자 일별 변동 예측모형의 주요 규칙은 <표 6>와 같다.

1번 규칙은 코스피 지수가 전일비 상승하였을 때 삼성전자 주가가 상승함을 의미한다. 2번 규칙은 코스피 지수가 전일비 하락하였을 때, 한국은행 기준금리가 1.25% 이하일 때 삼성전자 주가가 하락함을 의미한다. 3번 규칙은 코스피 지수가 전일비 하락하였을 때, 한국은행 기준금리가 1.25% 초과일 때, 소비자 물가지수가 전월비 상승하였고, 외국인 키워드가 뉴스 기사로 노출되지 않았을 때 삼성전자 주가가 하락함을 의미한다. 4번 규칙은 코스피 지수가 전일비 하락하였고, 한국은행 기준금리가 1.25% 초과일 때, 소비자 물가지수가 전월비 하락하였고, 금리 키워드가 뉴스 기사로 노출되었을 때 삼성전자 주가가 하락함을 의미한다.



[그림 3] 삼성전자 일별 변동 의사결정나무

<표 6>. 삼성전자 일별 변동 주요 규칙

구분	규칙 설명	규칙 (전일비)	예측수	적중률
1	KOSPI_daily = above	Y	331	75.8%
2	KOSPI_daily = under & KOR_short_ir <= 1.25	N	77	92.2%
3	KOSPI_daily = under & KOR_short_ir > 1.25 & CPI = above & word_foreign = N	N	46	91.3%
4	KOSPI_daily = under & KOR_short_ir > 1.25 & CPI = under & word_interest = Y	N	49	77.5%

삼성전자의 주요 규칙들을 통해 얻을 수 있는 시사점은 다음과 같다.

첫 번째, 삼성전자 주가 변동은 코스피 일별 변동과 밀접한 연관이 있는 것으로 판단된다. 두 번째, 일반적으로 금리 인하는 기대 심리에 의해 단기적으로 주가 상승에 영향을 미칠 수 있지만, 반드시 금리 인하가 주가 상승을 견인한다고 볼 수 없다.

연구 자료의 해당 기간 중에 한국은행 기준금리는 5회 변경되었고, 2019년 10월 16일 이후로 1.25% 이하로 금리가 변경됨을 감안할 때, 1.25%라는 기준은 시간적 의미를 내포한다. 즉, 2019년 10월 16일 이후에는 코스피 일별 변동 하락이 삼성전자 주가 하락과 밀접한 연관이 있는 것으로 파악된다.

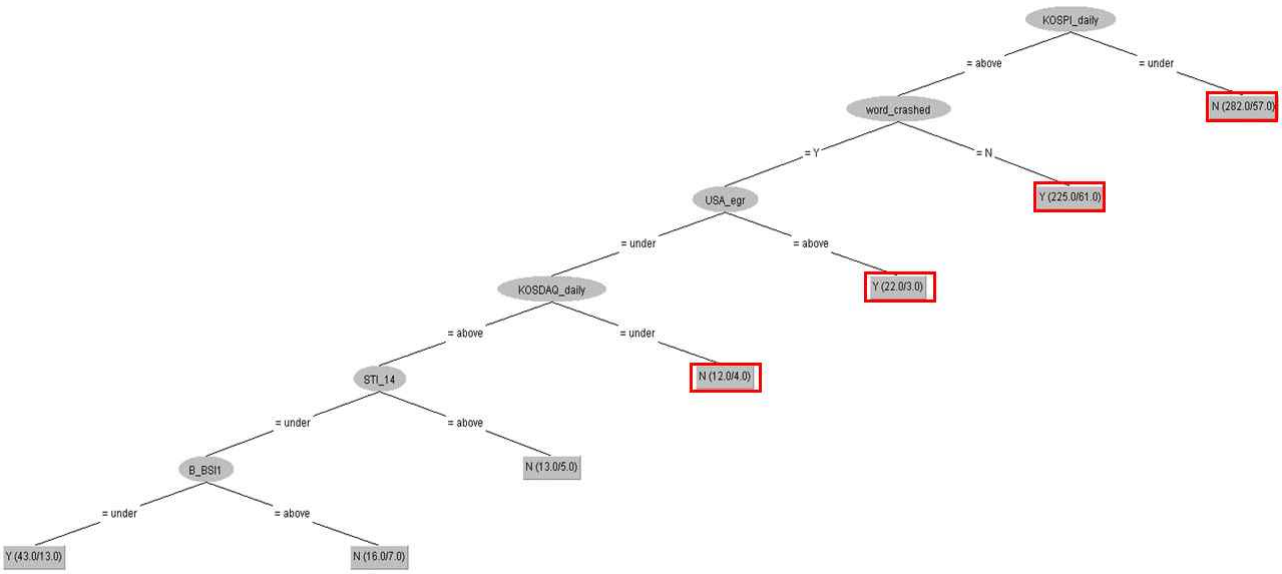
세 번째, 2019년 10월 16일 이전에 소비자물가 총 지수는 전월비 상승과 하락을 반복하였는데, 이때 해당되는 키워드가 조합되어 삼성전자 주가 상승 및 하락과 연관이 있

는 것으로 보인다. 특히 삼성전자의 외인 보유량이 평균적으로 50% 이상을 상회하는데, 외국인이 뉴스 기사에 노출되지 않았을 때 주가가 높은 확률로 하락한 것으로 파악된다.

다. 미래에셋대우 일별 변동 예측모형

미래에셋대우 일별 변동 예측모형의 분석 결과인 나무모형은 [그림 4]와 같다.

[그림 4]는 미래에셋대우 일별 변동 의사결정나무이다. 다양한 변수들이 결정요인으로 등장하였으며, 주요 규칙은 빨간색 상자로 표시하였다. 이를 토대로 도출한 미래에셋대우 일별 변동 예측모형의 주요 규칙은 <표 7>과 같다.



[그림 4] 미래에셋대우 일별 변동 의사결정나무

<표 7>. 미래에셋대우 일별 변동 주요 규칙

구분	규칙 설명	규칙 (전일비)	예측수	적중률
1	KOSPI_daily = under	N	282	79.7%
2	KOSPI_daily = above & word_crashed = N	Y	225	72.8%
3	KOSPI_daily = above & word_crashed = Y & USA_egr = above	Y	22	86.3%
4	KOSPI_daily = above & word_crashed = Y & USA_egr = under & KOSDAQ_daily = under	N	12	66.6%

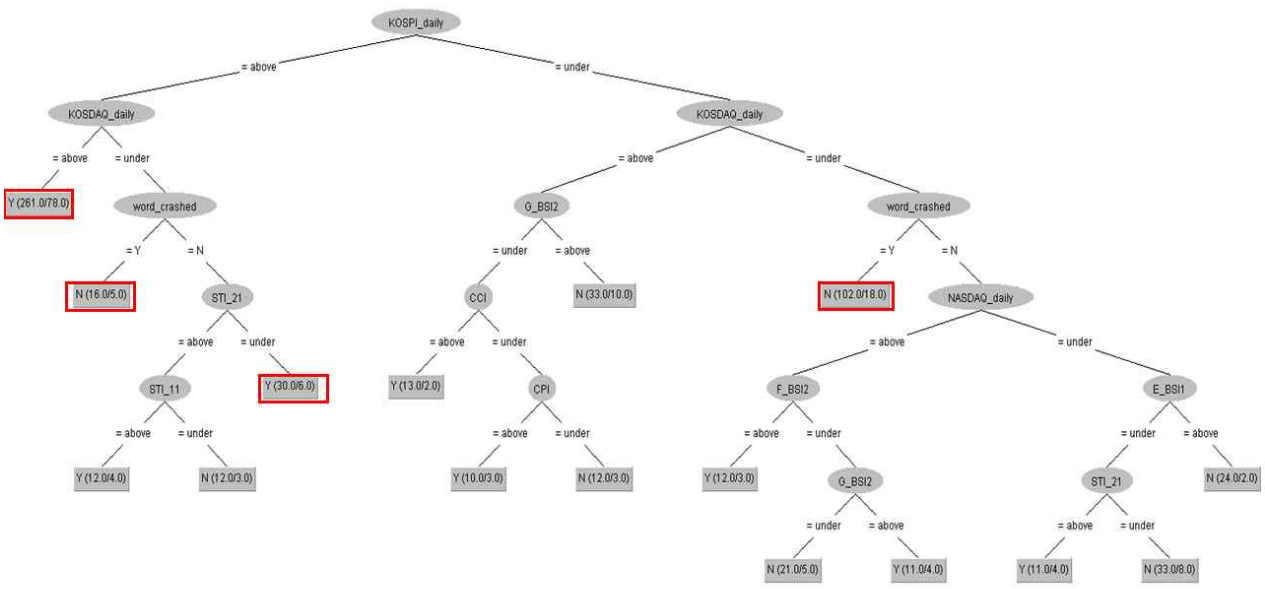
1번 규칙은 코스피 지수가 전일비 하락하였을 때 미래에셋대우 주가가 하락함을 의미한다. 2번 규칙은 코스피 지수가 전일비 상승하였을 때, 급락 키워드가 뉴스 기사에 노출되지 않았을 때 미래에셋대우 주가가 상승함을 의미한다. 3번 규칙은 코스피 지수가 전일비 상승하였고, 급락 키워드가 뉴스 기사에 노출되었으며, 미국의 실질경제성장률이 전분기 대비 상승하였을 때 미래에셋대우 주가가 상승함을 의미한다. 4번 규칙은 코스피 지수가 전일비 상승하였고, 급락 키워드가 뉴스 기사에 노출되었으며, 미국의 실질경제성장률이 전분기 대비 하락하였고, 코스닥 지수가 하락하였을 때 미래에셋대우 주가가 하락함을 의미한다.

미래에셋대우의 주요 규칙들을 통해 얻을 수 있는 시사점은 다음과 같다.

첫 번째, 코스피 지수와 코스닥 지수가 하락하였을 때 미래에셋대우의 주가는 하락하는 것으로 파악된다. 두 번째, 코스피 지수가 상승하였을 때 급락 키워드가 뉴스 기사에 노출되지 않으면 주가가 상승함을 알 수 있다. 그러나 급락 키워드가 뉴스 기사에 노출되었을지라도 미국의 실질경제성장률이 전분기 대비 상승하였다면, 주가는 상승한다.

라. LG화학 일별 변동 예측모형

LG화학 일별 변동 예측모형의 분석 결과인 나무모형은 [그림 5]와 같다.



[그림 5] LG화학 일별 변동 의사결정나무

[그림 5]는 LG화학 일별 변동 의사결정나무이다. 다양한 변수들이 결정요인으로 등장하였으며, 주요 규칙은 빨간색 상자로 표시하였다. 이를 토대로 도출한 LG화학 일별 변동 예측모형의 주요 규칙은 <표 8>과 같다.

1번 규칙은 코스피 지수가 전일비 상승하였고, 코스닥 지수가 전일비 상승하였을 때 LG화학 주가가 상승함을 의미한다. 2번 규칙은 코스피 지수가 전일비 상승하였고, 코스닥 지수가 전일비 하락하였으며, 급락 키워드가 뉴스 기사에 노출되었을 때 LG화학 주가가 하락함을 의미한다. 3번 규칙은 코스피 지수가 전일비 상승하였고, 코스닥 지수가 전일비 하락하였으며, 급락 키워드가 뉴스 기사에 노출되지 않았고, 보험 주가지수가 전월비 하락하였을 때 LG화학 주가가 상승함을 의미한다. 4번 규칙은 코스피 지수가 전일비 하락하였고, 코스닥 지수가 전일비 하락하였으며, 급락 키워드가 뉴스 기사에 노출되었을 때 LG화학 주가가 하락함을 의미한다.

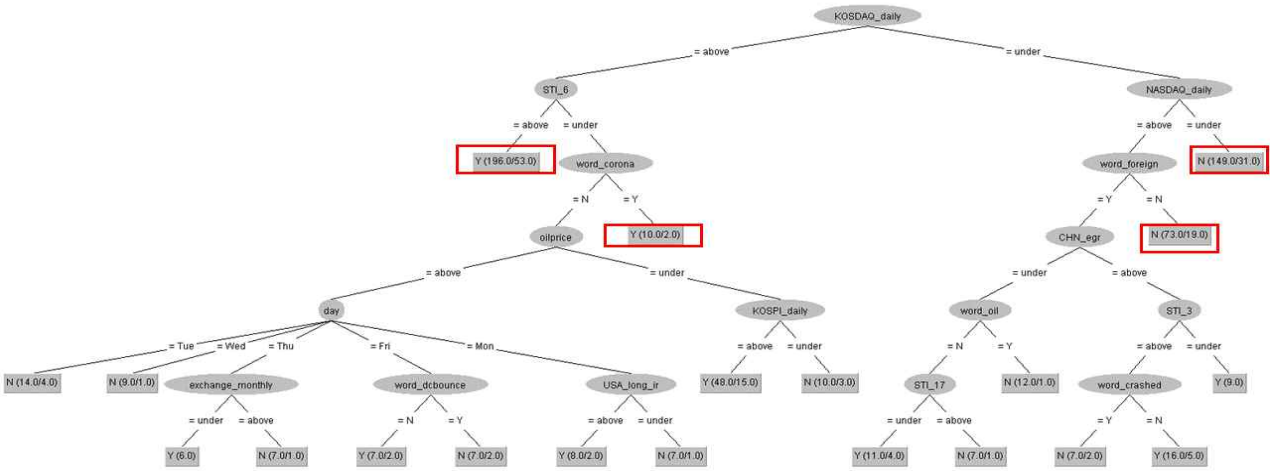
<표 8>. LG화학 일별 변동 주요 규칙

구분	규칙 설명	규칙 (전일비)	예측수	적중률
1	KOSPI_daily = above & KOSDAQ_daily = above	Y	261	70.1%
2	KOSPI_daily = above & KOSDAQ_daily = under & word_crashed = Y	N	16	68.7%
3	KOSPI_daily = above & KOSDAQ_daily = under & word_crashed = N & STI_21 = under	Y	30	80.0%
4	KOSPI_daily = under & KOSDAQ_daily = under & word_crashed = Y	N	102	82.3%

LG화학의 주요 규칙들을 통해 얻을 수 있는 시사점은 다음과 같다. 첫 번째, 전반적으로 LG화학 주가 상승과 하락이 코스피 지수 변동, 코스닥 지수 변동, 그리고 급락 키워드의 뉴스 기사 노출 여부에 따라 결정됨을 알 수 있다. 특히 코스피, 코스닥 지수가 하락하고 급락 키워드가 뉴스 기사에 노출된 경우 높은 확률로 주가가 하락함을 알 수 있다. 두 번째, 보험 주가지수의 월별 변동이 분지 기준으로 나타난 것에 대한 해석은 과대해석이 될 수 있으나, 동일 업종의 주가지수가 아님에도 불구하고 의사결정 상에 고려할 수 있는 변수가 될 수 있음을 시사한다.

마. 셀트리온 일별 변동 예측모형

셀트리온 일별 변동 예측모형의 분석 결과인 나무모형은 [그림 6]과 같다.



[그림 6] 셀트리온 일별 변동 의사결정나무

[그림 6]은 셀트리온 일별 변동 의사결정나무이다. 다양한 변수들이 결정요인으로 등장하였으며, 주요 규칙은 빨간색 상자로 표시하였다. 이를 토대로 도출한 셀트리온 일별 변동 예측모형의 주요 규칙은 <표 9>와 같다.

<표 9>. 셀트리온 일별 변동 주요 규칙

구분	규칙 설명	규칙 (전일비)	예측수	적중률
1	KOSDAQ_daily = above & STI_6 = above	Y	196	72.9%
2	KOSDAQ_daily = above & STI_6 = under & word_corona = Y	Y	10	80.0%
3	KOSDAQ_daily = under & NASDAQ_daily = under	N	149	79.1%
4	KOSDAQ_daily = under & NASDAQ_daily = above & word_foreign = N	N	73	73.9%

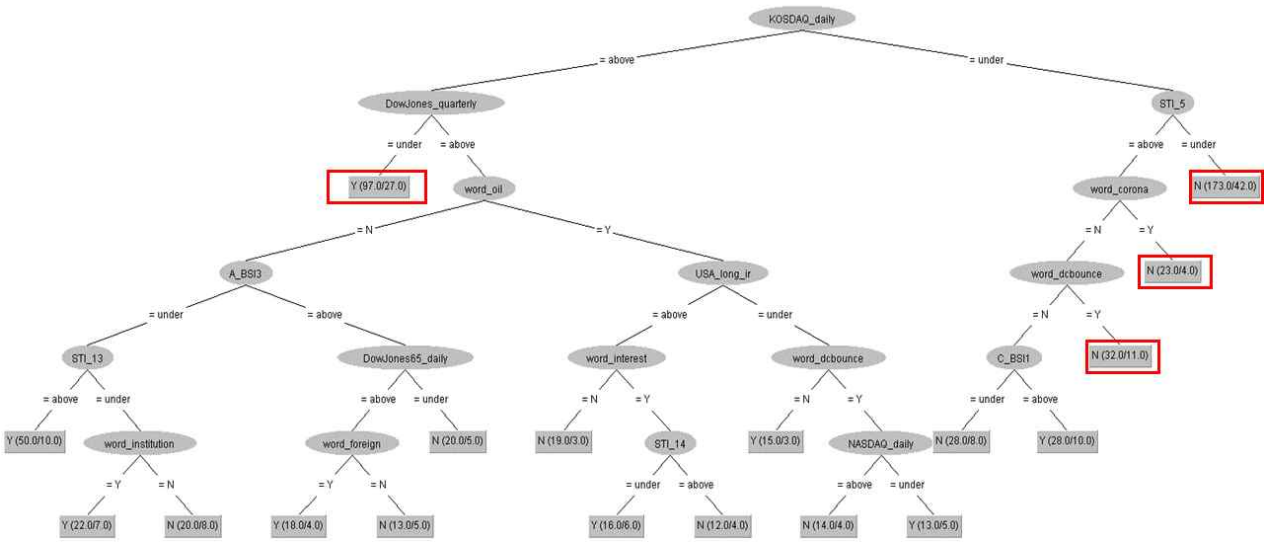
1번 규칙은 코스닥 지수가 상승하였고, 의약품 주가지수가 전월비 상승하였을 때 셀트리온 주가가 상승함을 의미한다. 2번 규칙은 코스닥 지수가 상승하였고, 의약품 주가지수가 전월비 하락하였으며, 코로나 키워드가 뉴스 기사에 노출되었을 때 셀트리온 주가가 상승함을 의미한다. 3번 규칙은 코스닥 지수가 하락하였고, 나스닥 지수가 하락하였을 때 셀트리온 주가가 하락함을 의미한다. 4번 규칙은 코스닥 지수가 하락하였고, 나스닥 지수가 상승하였으며, 외국인 키워드가 뉴스 기사에 노출되지 않았을 때 셀트리온 주가는 하락함을 의미한다.

셀트리온의 주요 규칙들을 통해 얻을 수 있는 시사점은 다음과 같다. 첫 번째, 셀트리온은 코스닥 지수와 나스닥 지수의 변동에 따라 상승 혹은 하락을 하는 것으로 판단된

다. 또한 유사 업종인 의약품 주가지수의 상승에 따라 높은 확률로 주가가 상승한 것으로 해석된다. 두 번째, 코로나 키워드가 뉴스에 노출되었을 때 셀트리온의 주가가 상승한 규칙을 해석해보면, 코로나19에 대한 치료제 및 의약품 개발 등의 기대심리가 반영된 것으로 판단된다. 이는 코로나19로 인한 대표적인 수혜주로 바이오 및 제약주가 부각되고 있는 현 상황을 반영하는 규칙이라 할 수 있다.

바. 하나투어 일별 변동 예측모형

하나투어 일별 변동 예측모형의 분석 결과인 나무모형은 [그림 7]과 같다.



[그림 7] 하나투어 일별 변동 의사결정나무

[그림 7]은 하나투어 일별 변동 의사결정나무이다. 다양한 변수들이 결정요인으로 등장하였으며, 주요 규칙은 빨간색 상자로 표시하였다. 이를 토대로 도출한 하나투어 일별 변동 예측모형의 주요 규칙은 <표 10>과 같다.

1번 규칙은 코스닥 지수가 전일비 상승하였고, 다우존스 지수가 전분기비 하락하였을 때 하나투어 주가가 상승함을 의미한다. 2번 규칙은 코스닥 지수가 전일비 하락하였고, 화학 주가지수가 전월비 하락하였을 때 하나투어 주가는 하락함을 의미한다. 3번 규칙은 코스닥 지수가 전일비 하락하였고, 화학 주가지수가 전월비 상승하였으며, 코로나 키워드가 뉴스 기사에 노출되었을 때 하나투어 주가는 하락함을 의미한다. 4번 규칙은 코스닥 지수가 전일비 하락하였고, 화학 주가지수가 전월비 상승하였으며, 코로나 키워

드가 뉴스 기사에 노출되지 않았고, 반등 키워드가 뉴스 기사에 노출되었을 때 하나투어 주가는 하락함을 의미한다.

<표 10>. 하나투어 일별 변동 주요 규칙

구분	규칙 설명	규칙 (전일비)	예측수	적중률
1	KOSDAQ_daily = above & DowJones_quarterly = under	Y	97	72.1%
2	KOSDAQ_daily = under & STI_5 = under	N	173	75.7%
3	KOSDAQ_daily = under & STI_5 = above & word_corona = Y	N	23	82.6%
4	KOSDAQ_daily = under & STI_5 = above & word_corona = N & word_dcbounce = Y	N	32	65.6%

하나투어의 주요 규칙들을 통해 얻을 수 있는 시사점은 다음과 같다. 첫 번째, 코스닥 지수 변동이 뿌리 노드에 등장하였고, 코로나 키워드가 뉴스 기사에 노출된 경우 하나투어의 주가가 높은 확률로 하락하였다. 이는 실제 코로나19로 인한 관광업에 대한 타격이 실재함으로 해석된다. 두 번째, 다우존스 지수와 화학 주가지수 등 거시경제지표가 분지의 기준으로 등장하였다. 그 중에서도 화학 주가지수에 대한 해석은 다음과 같다.

화학 주가지수와 코로나 키워드의 조합은 유추하건대 코로나19 이후의 전기차 관련 수혜주들이 화학 주가지수 범주 내에 포함된 것으로 보이며, 동일 기간에 전기차 관련 수혜주의 주가지수가 상승하였고, 마찬가지로 코로나19 키워드가 뉴스 기사에 많이 노출되었다. 그러나 무리한 해석일 가능성이 있기 때문에, 예측수와 적중률이 준수함에도 불구하고 일반화의 어려움은 존재한다.

Ⅴ. 결론

본 연구는 Istans에서 제공하는 거시경제지표와 Dataguide 5.0의 개별 기업의 종가 및 일별 국내·외 주가지수, 네이버 증권의 메인 뉴스 상에 노출된 기사를 키워드 별로 크롤링하여 데이터웨어하우스를 구축함으로, 코스피 일별 변동 예측모형과 개별 기업 주식 변동 예측모형을 개발하였다.

분석 기법으로 데이터 마이닝의 의사결정나무 분석을 시행하였으며, 16개의 예측모형을 개발하였고, 예측률 및 정분류율에 대한 성능 평가를 진행하였다. 최종적으로 6개의 예측모형을 선정하였으며, 선정된 예측모형의 나무모형을 각각 해석하고 주요 규칙들을 도출하였다.

결과적으로 코스피 일별 변동 예측모형이 가장 높은 예측률을 기록하였으나, 개별 기

업의 일별 변동에 대한 규칙이 주로 도출되었으므로 해석 상의 어려움이 존재하였다. 또한 개별 기업의 주가 일별 변동 예측모형에 대한 해석은 다양한 거시경제지표와 키워드 변수 등이 분지 기준으로 등장하였으나, 코스피와 코스닥 지수의 변동에 따른 주가 변동 여부가 주로 구분되었다.

본 연구의 한계점은 다음과 같다. 개별 주식의 거래량, 수익률, 재무지표 등을 활용하지 않았다. 또한 단일 알고리즘을 통한 분석을 시행하였기 때문에 다차원적인 분석이 진행되지 않았다. 이를 향후 연구를 통해 보완하고자 한다.

참고문헌

고강석(2018). “업종별 주가지수와 주가-환율 관계의 안정성”, 한국자료분석학회, 20(2), pp. 815-828.
김용재, 이상춘(2017), “한국의 거시경제 지표가 기업의 주가수익률에 미치는 영향분석: SVAR 활용”, 전문경영인연구, 20(3), pp. 281-303.
김유신, 김남규, 정승렬(2012), “뉴스와 주가; 빅데이터 감성분석을 통한 지능형 투자의 사결정모형”, 지능정보연구, 18(2), pp. 143-156.
김주일(2013), “KOSPI지수 및 KOSDAQ지수와 환율과의 상호연관성에 관한 연구”, 한국산업경제학회 정기학술발표대회 초록집, pp.37-54.
박은정, 조성준(2014), “KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지”, 한글 및 한국어 정보처리 학술대회 논문집, 26, pp. 133-136.
박정미, 박성태(2019), “ICT 관련 상호변경이 주가에 미치는 영향”, 대한경영학회지, 32(6), pp. 1019-1040.