「정보시스템연구」 제30권 제3호
한국정보시스템학회
2021년 9월, pp. 21~45

http://dx.doi.org/10.5859/KAIS.2021.30.3.21

데이터 마이닝 기법을 통한 COVID-19 팬데믹의 국내 주가 영향 분석: 헬스케어산업을 중심으로*

김덕현** · 유동희*** · 정대율****

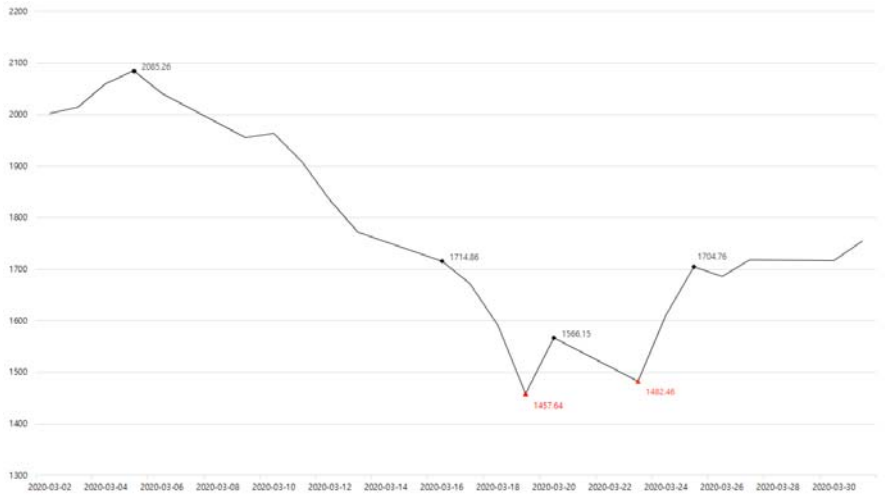
〈목 차〉

Ⅰ. 서 론

2020년도 한국의 유가증권시장(이하 주식시장)은 COVID-19(이하, 코로나19)로 인한 사상 유례없는 팬데믹 현상으로 매우 큰 변화를 경험하였다. 이로 인하여 실물경제와 금리는 급속히 하락하였으며, KOSPI(KOrea composite Stock Price Index, 이하 코스피)에도 엄청난 영향을 미쳤다. 국내 주식시장은 경제적으로 큰 사건이 있을 때마다 크게 요동치는 모습을 보였다(이용대 등, 2015). 실제로 국내 주식시장은 2020년 3월에 매우 큰 변화의 폭을 경험하였는데, 이는 국내·외 코로나19의 확산과 세계보건기구(이하, WHO)의 팬데믹 선언 등과 맞물려 일어났다.

KOSIS(2020)에 따르면, 국내 코로나19 확진자의 수는 2020년 3월 1일 3,736명에서 2020년 3월 31일 9,786명으로 급증하였다. 국외에서도 마찬가지로 코로나19의 확진세가 두드러지면서 WHO(2020)는 2020년 3월 13일에 코로나19에 대한 팬데믹 선언을 하였다. <그림 1>은 코

* 이 논문은 2021년 경영정보학 관련 춘계통합학술대회에서 정보시스템학회 최우수상으로 선정된 논문을 수정·보완하였다.
** 경상국립대학교 경영정보학과, kdh294@gnu.ac.kr (주저자)
*** 경상국립대학교 경영정보학과, dhyoo@gnu.ac.kr (공저자)
**** 경상국립대학교 경영정보학과, dyjeong@gnu.ac.kr (교신저자)



<그림 1> 2020년 3월 KOSPI 종가 변동 추이

로나19의 파급이 가장 크게 나타난 2020년 3월 KOSPI 종가의 일별 변동 추이를 나타낸 것이다. 국내의 코로나19 확산세와 더불어 팬데믹 선언 전·후로 매우 급변하는 양상을 살펴볼 수 있다.

전반적으로 월초 2,000pt 선을 유지하다가 크게 하락하였고, 다시 반등하는 추세를 보인다. 특히 3월 19일에 1,457.64pt를 기록하여 역대 최고 당일 하락을 기록하였으며, 약 10일 만에 500pt가 하락하였다[1]. 그러나 짧은 기간 내에 빠른 속도로 이른바 'V자 반등'에 성공함으로 증시가 회복되었고, 이후 코로나19 백신 개발에 대한 기대 심리가 반영된 제약·바이오산업(이하, 헬스케어산업과 혼용되어 사용됨) 관련주의 주가가 큰 폭으로 상승하며 소위 말하는 코로나19로 인한 수혜주로 떠올랐다(김민정, 권택호, 2021).

본 연구에서 코로나19와 같은 글로벌 팬데믹 현상이 주식시장에 어떠한 영향을 미치는가를 분석하기 위하여 다음과 같은 연구 문제를 제기한다.

Q1. 코로나19의 영향으로 인한 전체 주식시장의 변화 패턴은 어떻게 나타났는가?

Q2. 코로나19의 영향으로 인한 헬스케어산업분야의 주가 변화 패턴은 어떻게 나타났는가?

Q3. 코로나19와 같은 팬데믹 현상에서 주가에 영향을 미치는 핵심 변수는 무엇인가?

본 연구에서는 코로나19로 인한 한국 주식시장의 변화 큰 흐름을 먼저 파악하고, 제약·바이오산업 중심의 헬스케어산업 분야의 주요 주식의 주가 변화를 분석하기 위하여, 주식시장과 정부기관에서 제공하는 다양한 원천의 DB로부터 분석에 필요한 데이터를 수집하였다. 그리고 증권 관련 뉴스가 시장과 개인에게 영향을 줄 수 있다는 점과 코로나19에 대한 정보가 매스

1) 2020년 3월 10일 1,962.27pt에서 동년 동월 19일 1,457.64pt로 하락, ▽505.29pt.

미디어 및 인터넷 뉴스를 통해 전달될 수 있다는 점을 고려하여, 코로나19와 관련된 키워드를 중심으로 증권 관련 뉴스에 게재된 기사들을 웹으로부터 크롤링하였다. 또한 코로나 감염 환자 수와 같은 코로나19 관련 정보는 통계청의 코로나19 발생 현황 자료를 활용하였다. 본 연구에서는 수집된 데이터를 크게 두 가지로 구분하여 실험을 진행하였다. 먼저, 연구자료의 기간에 따른 차이를 규명하기 위하여 데이터셋(data set)을 다르게 구성하고, 이들 자료에 대하여 데이터 마이닝 기법을 적용하여 분석을 시도하였다. 또한, 목표변수를 한국종합주가지수(KOSPI)와 KOSPI200 헬스케어로 구분하여 실험을 진행하였다.

Ⅱ. 이론적 기초 및 선행 연구

2.1 연구방법론 관련 선행 연구

데이터 마이닝이란 많은 양의 데이터로부터 의미 있는 규칙이나 패턴을 도출하는 기법을 의미한다. 본 연구에서는 데이터 마이닝 기법 중 의사결정나무를 통해 코로나19 팬데믹으로 인한 한국 증시의 영향과 주요 패턴을 발견하고자 하였다. 분석을 사용된 데이터는 여러 가지 원천으로부터 수집되어졌다. 먼저, 주식시장의 종목별 종가 공시 자료를 중심으로 주요 분석 자료를 수집하였으며, KOSIS(2020)에서 발표하는 일일 코로나19 발생 현황 자료를 활용하였다. 그리고 뉴스 기사를 크롤링하여 비정형적 데이터를 수집하였다.

각종 언론 매체의 뉴스 기사, SNS, 주식 토론방 등으로부터 수집되는 텍스트 데이터를 비정형 데이터라고 하는데, 이러한 비정형 텍스트들을 수집하여 변수로 추출·활용하는 연구들은 텍스트 마이닝에 관련된 연구라 할 수 있다. 본 연구에서는 뉴스 기사 내에 포함된 키워드를 도출하고, 증시의 개·폐장 여부를 고려하여 시간적 특성을 반영함으로 '헬스케어 분야 및 코로나19에 대한 키워드 변수'를 분석에 활용하였다.

의사결정나무 분석기법은 데이터마이닝 분야에서 가장 많이 사용되는 기법 중의 하나이다. 의사결정나무 분석기법은 데이터마이닝에서 탐색과 모형화라는 두 가지 특성을 모두 가지고 있다. 즉, 의사결정나무 분석기법은 모수적 모형을 분석하기 위해서 사전에 이상치를 검색하거나 분석에 필요한 변수 또는 모형에 포함되어야 할 상호작용의 효과를 찾아내기 위해서 사용될 수도 있고, 의사결정나무 자체가 분류 또는 예측모형으로 사용될 수도 있다. 따라서 의사결정나무 분석기법은 의사결정 규칙을 나무구조로 도표화하여 제시할 수 있어 분류와 예측을 수행하는 분석 방법으로 유용하다. 또한 예측의 과정이 나무구조에 의한 추론규칙에 의해서 표현되기 때문에, 그 과정을 쉽게 이해하고 추론과정을 쉽게 설명할 수 있다는 장점을 가진다(Song and Lu, 2015; 강현철 등, 2014; 김덕현 등, 2019).

주식시장에서 텍스트 마이닝 기법을 활용한 선행 연구들은 다음과 같다. 텍스트 마이닝 기법을 국내 주식시장에 적용한 연구에는 KoNLpy를 활용하여 뉴스와 주가에 대한 영향관계를 분석한 연구(김유신 등, 2012), SNS에 속하는 종목 토론방에 대해 OKT(Open Korea

Text)와 딥러닝 기법을 활용하여 코스피200의 20개 종목에 대한 주가 방향성을 예측한 연구(김명진 등, 2020), Krkwic를 활용하여 인터넷 주식 뉴스를 통한 종합주가지수 예측을 시도한 연구(유지돈, 이익선, 2018), SentiWordNet 기반의 뉴스 감성분석과 SVM 분석을 활용하여 다우존스 지수와 S&P500 지수를 예측한 연구(홍태호 등, 2017) 등이 있다. 이 밖에도 인터넷 검색량과 개별 기업의 주가변동성의 관계를 산업별로 분석한 전새미 등(2020)의 연구, R-N2H4 패키지를 활용하여 코로나19가 자산시장(주택시장, 주식시장)에 미치는 영향을 분석한 전해정(2020) 연구도 웹으로부터 데이터를 수집하고 이를 변수로 활용한 연구들이라 할 수 있다.

주식시장은 아니지만 자산관리 분야에 텍스트 마이닝 기법을 적용한 연구도 있다. 손애린 등(2020)은 국내 자산관리 핀테크 기업의 성공요인을 도출하기 위하여 텍스트 마이닝 기법을 이용하였다. 이들은 사용자 리뷰를 형태소 분석으로 전처리한 후 LDA 토픽모델링을 통해 주요 요소를 추출하고, 속성별 감성분석을 통하여 사용자들이 자산관리에 대하여 보이는 반응의 형태(긍정적 반응, 부정적 반응)를 분석하였다. 그리고 이것을 계수화하여 속성별 감성지수를 산출하였다.

2.2 구성변수 관련 선행 연구

주식시장의 주가 예측에 관하여 수많은 이론과 실증연구가 있다. 본 연구에서는 이 중에서 본 연구에서 직·간접적으로 활용된 변수들을 중심으로 관련 있는 선행 연구를 검토하고자 한다.

김정희와 장병기(2010)는 거시경제 및 해외변수들의 국내 주가에 대한 영향력을 파악하는 연구를 수행하였다. 주가에 영향을 미치는 주요 변수로 '환율', '유가', '해외주가', '장기금리의 3년 만기 회사채수익률', '단기금리의 콜금리', '두바이유 유가', 'S&P500' 등이 있다. 그들의 연구 결과에 의하면, 환율이 주가에 미치는 영향이 유의하며, 과거 양(+)의 영향에서 현재 음(-)의 영향으로 변화한 것으로 나타났다. 특히, 미국 주가의 영향력이 국내의 구조 변화(외환위기를 기점) 이후 더욱 강화되었다. 주가 결정에 있어서 실물적 요소가 중요하게 작용하던 외환위기 이전과 달리, 외환위기 이후에 금융적 요소가 중요하게 작용한다고 보고하고 있다. 이는 미국 증시와 국내 증시의 연관성을 입증한 부분이라 할 수 있다.

김주일(2013)은 코스피와 코스닥, 환율과의 상호연관성을 분석하였다. 주요 변수로 'KOSPI 지수', '코스닥 지수', '달러 환율' 등을 활용하였으며, 인과관계 분석과 충격반응 분석 및 분산분석을 실시하였다. 그 결과 외국인 투자자가 국내 주식을 매매하는 비중에 따라 국내 주가지수와 환율 간의 영향이 변동됨을 규명하였다. 또한, 이용대 등(2015)은 한국 증시에서 주식수익률과 거래량 간의 동적 관계를 규명하고자 하였으며, 주요 변수로 'KOSPI(종합주가지수)', '규모별 KOSPI 구분 지수(대형, 중형, 소형)과 해당 거래량', '투자 주체별 거래량(개인, 외국인, 기관)'을 선정하였다. 주요 결과로 과거에는 수익률이 거래량을 선도하는 모습을 보이다가, 금융위기 등의 금융시장에서의 사건 이후 시간이 지남에 따라 거래량과 주가 간의 상호 관계

가 약해진 것으로 나타났다. 이에 대하여 투자자들이 주식을 거래하고 주가를 평가하는 데 있어 과거에 비해 정보의 획득 방법이 수월해지고, 주식을 분석하는 방법이 발전하였기 때문으로 해석하고 있다.

김용재와 이상춘(2017)은 SVAR(구조적 벡터 자기회귀)을 활용하여 한국의 거시경제지표가 기업의 주가수익률에 미치는 영향을 분석하였다. 이들이 사용한 주요 변수로는 '환차익(손)율', '원유가', '총통화', '3년 만기 국고채수익률', '국제수지' 등의 거시경제변수가 있으며, 이를 통해 개별 기업의 주가수익률을 분석하였다. 주요 결과로 국제수지가 가장 큰 영향을 나타냈으며, 이는 한국이 수출주도형 산업으로 외환의 주 공급처가 기업임을 확인할 수 있는 결과라고 해석하고 있다. 한편, 고강석(2018)은 업종별 주가지수와 주가, 환율 관계의 안정성을 분석하였다. 주요 변수로 '업종별 주가지수', '환율' 등이 있으며, 전통적 접근과 포트폴리오 접근의 두 가지 방법에 의해 결과를 해석하였다. 주요 결과로 주가와 환율의 동태적 관계는 과거에 비하여 현재에는 미비한 것으로 견해를 제시하였다.

고승의(2018)는 한국 주식시장에서 성장주와 가치주의 초과수익률을 비교한 연구를 제시하였는데, 이때 제약·바이오 주를 IT 관련주와 함께 성장주로 구분하였다. 2001년 이후의 16년 동안의 기간에 대해 분석하였으며, 주요 결과로 성장주가 가치주보다 높은 투자수익률을 보인 기간은 3개년에 불과하며 특정 시점에서 성장주의 수익률이 높은 구간이 있으나, 그 기간이 짧고 성장주 중에서도 몇 개의 특정한 성장주만이 과도하게 높게 수익률을 기록한 것으로 나타났다. 이에 대하여 우리나라 증권시장에서 효율적 시장가설의 실증적 타당성이 매우 제한적임을 시사하였다.

성노윤과 남기환(2018)은 산업군 내 동질성을 고려하여 온라인 뉴스 기반에 주가 예측을 시도하였다. 이를 위해 '국제산업표준체계에 의한 구분'과 '기계학습에 의한 군집 구분'의 성능을 비교하였고, 분석기법으로 K-means 군집분석과 다중커널학습의 EasyMKL를 활용하였다. 주요 결과 중의 하나로 제약의 경우 음식료업과 함께 동질 산업군 내에서 이질성이 작았으며 국제산업표준체계가 더 나은 결과를 보였고, 소재업의 경우 이질성이 크기 때문에 군집 구분에 의한 구분이 더 나은 결과를 보였다.

또한, 전해정(2020)은 코로나19가 주식시장과 부동산시장에 미친 영향에 관하여 베이지언 네트워크 분석을 하였는데, 주요 변수로 '코로나19 뉴스 키워드 변수', '코로나19 신규확진자', '경기도 주택매매량', 'KOSPI 수익률' 등을 활용하였다. 주요 결과로 코로나19 신규 확진자 수와 코로나19 뉴스 기사 간의 관계가 정(+)의 방향 관계에 있으나, 이 밖의 결과들은 코로나19가 자산시장에 미치는 영향력을 온전히 판단하기 힘든 것으로 요약하였다.

김민정과 권택호(2021)는 코로나19 상황에서의 제약기업 시장성과의 변화, 그리고 기업의 특성과의 관련성에 대하여 연구하였는데, 주요 결과로 최고경영자의 학력만이 시장성과의 변화에 유의적인 것으로 확인되었다. 반면, 연구개발이나 수출 등과 같은 기업의 역량에 대한 변수들은 유의미하지 않은 것으로 나타났다. 이에 대하여 코로나19 상황에서 한국 제약·바이오 기업의 높은 시장성과는 기업의 축적된 역

량의 결과이기보다는 산업의 특성이 반영된 결과라고 해석하고 있다.

지금까지 진행된 선행 연구들을 종합해보면 주가 혹은 종합주가지수를 예측하기 위하여, 기존의 재무적 지표(주가수익비율 등)를 활용한 연구들의 한계를 개선하는 방법론적 시도로 거시경제변수, 공표된 자료, 뉴스 기사 등의 다양한 변수를 고려하고 있는 것으로 보인다.

Ⅲ. 연구 방법

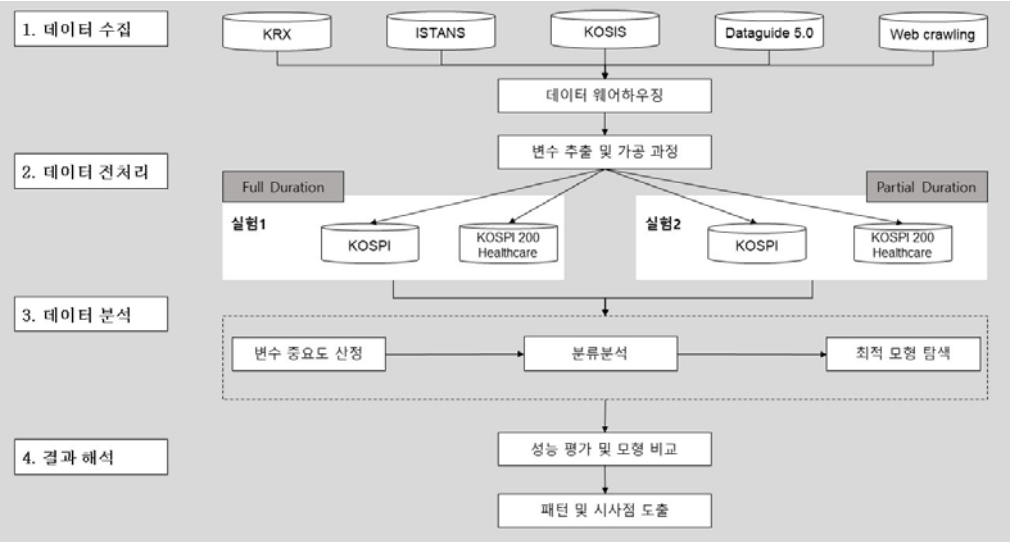
3.1 연구 프레임워크

본 연구는 코로나19로 인한 팬데믹 현상이 헬스케어산업을 중심으로 한 한국의 증권시장에 미치는 영향을 분석하는 것이 목적이다. 이를 위해 앞의 2장에서 살펴본 선행 연구에 기초하여 다양한 원천의 빅데이터를 활용한 탐색적 접근법을 택하였다. 본 연구의 전체 진행 과정은 <그림 2>와 같이 크게 4단계로 구분된다.

첫째, 데이터 수집 단계에서는 각각의 데이터 원천으로부터 데이터를 수집하고 데이터 웨어하우스를 구축하였다. 데이터 웨어하우스는 시스템 등에 분산되고 단절되어 있는 정보를 추출·변환·통합의 과정을 거쳐 저장·축적함으로써, 사용자가 원하는 분석정보를 제공하는 환경을 의미한다(정용완, 정승국, 2012).

둘째, 데이터 전처리 단계에서는 수집된 데이터를 분석에 알맞게 가공하고, 각 실험에 맞게 변수(feature)와 인스턴스(instance)를 추출하고 선별하였다. 실험1과 실험2의 차이는 연구 자료의 기간에 따른 차이이며, 전자는 수집된 자료의 전체 기간(3개년)에 해당되고 후자는 2020년도 1개년에 해당된다. 개별 데이터셋(dataset)은 목표변수에 따라 구분되며, 각각의 데이터셋은 목표변수 1개와 독립변수 48개로 구성된다.

셋째, 데이터 분석 단계는 이득비에 근거한



<그림 2> 연구 프레임워크

변수의 중요도를 산정하고 분류분석을 시행한 후 모형의 성능을 평가한다. 그 후 중요도가 낮은 변수를 제거한 후 다시 분류분석을 시행하는 과정들을 반복 수행한다. 이러한 반복 작업은 최적의 모형을 탐색했을 때 완료되며 다음 단계로 넘어간다.

넷째, 결과 해석 단계에서는 선정된 모형에 대한 성능 평가 및 비교를 실시하고, 의사결정나무를 기반으로 패턴 및 시사점 도출한다.

3.2 연구자료 수집

본 연구에서는 한국증권시장의 KOSPI와 KOSPI 200(Korea Stock Price Index 200, 이하 코스피200)에서 제공된 자료를 기초로 한다. KOSPI는 한국증권거래소의 유가증권시장에 상장된 회사들의 주식에 대한 총합인 시가총액의 기준시점(1980년 1월 4일)과 비교 시점을 비교하여 나타낸 지표다. 원래 명칭은 종합주가지수였으나, 2005년 11월 1일부터 지금의 명칭으로 바뀌어 사용되고 있다(한국증권거래소, 2021a). 반면, KOSDAQ(KOrea Securities Dealers Automated Quotation, 이하 코스닥)은 1996년 설립된 대한민국의 주식시장이다. 첨단 기술주 중심인 나스닥(NASDAQ) 시장을 본떠 만든 것으로, 유가증권 시장과는 규제 조치가 별개로 이루어지는 시장이다.

KOSPI 200은 1996년 5월 3일 시작된 코스피200 주가지수 선물 거래와 1997년 7월 7일 시작된 코스피200 주가지수 옵션거래를 위해 한국증권거래소에서 산출하여 발표하기 시작한 주가지수이며, 기준시점은 1990년 1월 3일이다. 이 지수는 시가총액이 크고 거래량이 많아 코스피를 대표할 수 있다고 간주 되는 종목 200개를 선정하여, 그들의 주가로부터 산출한다(한국증권거래소, 2021b). 코스피200은 코스피200 주가지수선물의 현물이며, 선물 및 옵션 거래의 기초가 되는 수치로 활용된다. 코스피200의 하위 구성은 '건설', '금융', '산업재', '중공업' 등이 있는데, 본 연구에서는 제약·바이오 산업에 해당하는 코스피200 헬스케어에 초점을 두고자 한다. 코스피200 헬스케어에 포함되는 개별 종목은 '녹십자', '삼성바이오로직스', '셀트리온', '유한양행' 등으로 국내를 대표하는 굴지의 제약·바이오 기업들이 대부분 포함된다.

본 연구는 2018년 1월 1일부터 2020년 12월 31일까지의 기간을 분석 자료의 전체 범위(실험1)로 지정하고, 2020년 1월 1일부터 2020년 12월 31일까지의 기간을 부분 범위(실험2)로 지정하였다. 한국 주식시장의 개장 여부를 기준으로, 휴장일을 제외하였다. 실험1의 샘플 수는 1,096일에서 휴장일 제외하여 738개이며, 실험2의 샘플 수는 366일에서 휴장일 제외하여 248개이다. 데이터는 한국거래소(KRX), 산업연구원(ISTANS), 통계청(KOSIS), Dataguide 5.0, investing.com, Naver 등으로부터 수집하였고, 선행 연구를 기반으로 주요 변수를 추출하였다.

3.3 연구변수 특성 분석

3.3.1 목표변수 특성 분석

본 연구의 목표변수로 자료 기간 내에 한국 주식시장의 전체적인 흐름을 파악하기 위한 코스피의 헬스케어 산업 분야를 대표할 수 있는 코스피200 헬스케어를 선정하였다. 목표변수

<표 1> 목표변수의 기술통계량과 하위집단 분포

구 분	Test 1 (n=738)		Test 2 (n=248)	
	KOSPI	KSP200 Healthcare	KOSPI	KSP200 Healthcare
Mean	2,216.76	2,184.05	2,220.07	2,446.72
Std.	208.41	420.51	275.32	495.16
Min.	1,457.64	1,412.55	1,457.64	1,427.96
Max.	2,873.47	3,377.43	2,873.47	3,377.43
Up	414	379	152	128
Down	324	359	96	120

모두 일별 종가 지수로 데이터의 유형은 수치형(numeric)이다. 데이터 전처리를 통해 일별 종가지수의 상승과 하락이라는 변동의 의미를 포함하면서 분석에 적합한 형태로 변환하였는데, 전일 대비 초과/이하로 범주화하였다. 즉, 목표변수의 데이터 유형을 수치형에서 명목형(nominal)으로 변경하였다. 코스피와 코스피200 헬스케어의 기술통계량은 <표 1>과 같다. 원자료(raw data)의 기술통계량과 데이터 전처리를 통한 목표변수 내 하위집단의 분포를 정리하였다. 실험1과 실험2의 범위가 다르기 때문에, 실험 구분에 따른 차이도 정리하였다.

<표 1>에 나타난 실험2의 결과를 살펴보면, 2020년도의 변화폭이 상당히 컸음을 알 수 있다. 이에 대한 근거로 실험2의 표준편차가 코스피와 코스피200 헬스케어 모두 실험1의 표준편차보다 크다는 점과 최솟값과 최댓값 모두 실험 구분에 무관하게 같다는 점을 들 수 있다. 김명진 등(2020)은 2019년 11월에서 2020년 7월까지의 기간이 국내 주식시장의 역사상 가장 많은 개인 투자 자금이 유입된 시기인 강세장에 해당한다고 평가한 바 있다.

앞서 언급한 바와 같이 연구자료의 기간에 따라 실험이 구분되는데, 실험1은 2018년 1월부터 2020년 12월까지, 실험2는 2020년 1월부터 12월까지 해당된다. 3개년의 최솟값과 최댓값, 1개년의 최솟값과 최댓값이 같다는 점과 표준편차도 약 70pt가량 차이가 있다는 의미는 '20년도 한 해 동안 한국 증시 변동 폭이 매우 컸다'라고 해석할 수 있다.

끝으로 Up과 Down은 각 지수의 값이 전날 종가에 비해 올랐는지, 같거나 내려갔는지(전일 대비 초과/이하)를 나타낸 목표변수의 하위집단을 의미한다. 전반적으로 초과(상승)가 이하(하락)보다 더 많은 횟수를 기록하였음을 알 수 있다.

3.3.2 독립변수 특성 분석

본 연구의 독립변수는 자료 수집 방법에 따라 크게 두 가지로 구분된다. 국내 기관들의 웹사이트로부터 공표된 자료를 내려받아 활용하거나, 텍스트 마이닝을 사용하여 키워드별 자료를 수집하였다. 독립변수의 기술통계량은 <표 2>와 같다. <표 2>는 분석자료의 전체 범위에 해당된다. 또한 비슷한 영역에 속하는 변수들은 음영의 어둡기를 동일하게 표현하여 구분하였다.

먼저, 기간이나 일시 등의 작용을 살펴보기

<표 2> 독립변수의 기술통계량

변수명	Mean	Std.	Min.	Max.
Date	해당 일자(예, 2018-01-02)			
YMW	해당 일자로부터 파생, 년월주(예, 2018Y1M1W)			
YQ	해당 일자로부터 파생, 년분기(예, 2018Y1st)			
USA	미국 증시 개장 여부			
Interest	해당시기 한국은행 기준금리			
Interest change	한국은행 기준 금리 변경 여부			
KOSPI	2216.76pt	208.41pt	1457.64pt	2873.47pt
kospi_trans	589,331,389주	283,699,418주	206,497,608주	1,984,181,619주
KOSDAQ	747.25pt	98.46pt	428.35pt	968.42pt
kosdaq_trans	1,085,279,446주	486,350,326주	421,570,139주	3,185,108,496주
KOSPI200 Healthcare	2,184.05pt	420.51pt	1,412.55pt	3,377.43pt
ksp200H_trans	9,148,012주	13,343,363주	1,290,447주	142,511,135주
KOSDAQ150 Healthcare	3,962.61pt	945.73pt	2,159.36pt	6,114.83pt
ksq150H_trans	23,124,871주	18,052,948주	4,708,539주	120,572,022주
SNP 500	2,959.58pt	289,42pt	2,237,40pt	3,735.36pt
DowJones	26,107.07pt	1,840.02pt	18,591.93pt	30,409.56pt
NASDAQ	8,524.76pt	1,517.37pt	6,192.92pt	12,899.42pt
Dollar_index	95.61	2.90	88.67	102.99
new_confirm	54.14명	157.51명	0명	1,090명
agg_confirm	5,437.49명	10,533.49명	0명	59,773명
new_death	0.84명	2.84명	0명	40명
agg_death	94.89명	172.29명	0명	879명
Corona	4.36회	8.92회	0회	55회
Phama	0.60회	1.31회	0회	12회
Bio	1.92회	4.01회	0회	51회
Vaccine	0.66회	2.73회	0회	32회
Virus	1.17회	2.81회	0회	25회
Confirm	0.30회	1.16회	0회	15회
Accumulate	0.06회	0.33회	0회	5회
Death	0.07회	0.40회	0회	4회
New	0.30회	0.71회	0회	5회

위해 해당 일자(Date)와 더불어, 이로부터 파생한 '년월주(YMW), 년분기(YQ)'를 분석에 투입하였다. 한국의 증시는 미국 증시와 밀접하다고 판단하여(김정희, 장병기, 2010), 미국 증시 개장일을 범주화(개장여부, USA)하였다. 거시경제변수로써 해당 시기의 한국은행 기준금리(Interest)를 고려하였고, 기준금리의 변경일을 바탕으로 변경 여부를 판단할 수 있는 변수(Interest change)를 파생하였다.

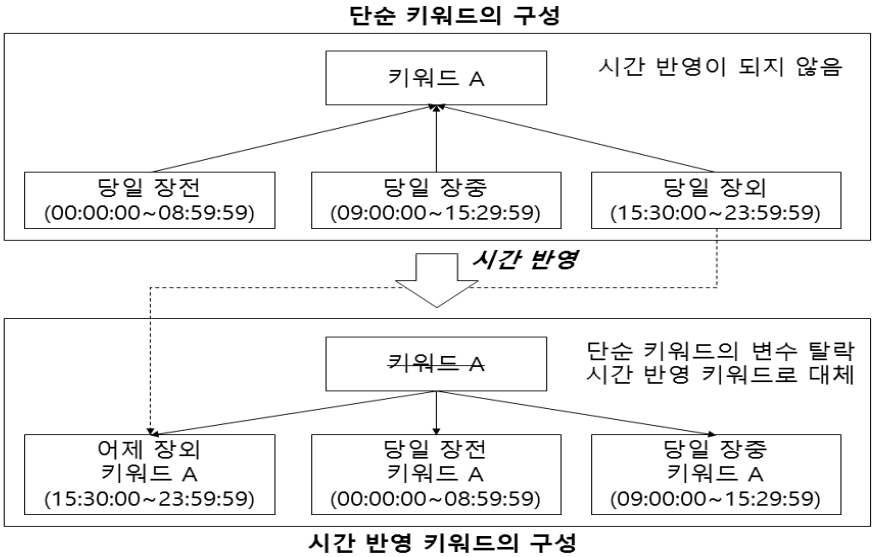
한국 증시를 나타내는 지표인 코스피(KOSPI)와 코스닥(KOSDAQ), 제약·바이오 분야 지표인 코스피200 헬스케어(KOSPI200_Healthcare), 코스닥 헬스케어를 전일 대비 일별 변동(초과/이하)으로 범주화하여 분석에 투입하였다. 이때 코스피와 코스피200 헬스케어는 각 데이터셋에 따라 다르게 투입하였는데, 목표변수가 코스피 일별 변동일 때 독립변수로 코스피를 제외하였고, 목표변수가 코스피200 헬스케어 일

별 변동일 때 독립변수로 코스피200 헬스케어를 제외하고 분석을 시행하였다. 변동성의 측면을 고려하여 코스피 거래량(kospi_trans), 코스닥 거래량(kosdaq_trans), 코스피200 헬스케어 거래량(ksp200H_trans), 코스닥150 거래량(ksq150H_trans)을 분석에 투입하였으며, 독립변수들의 단위 차이가 크게 나기 때문에 정규화(normalize)를 하였다.

미국 증시를 대표하는 S&P500(SNP500), 다우존스(DawJones), 나스닥(NASDAQ)도 마찬가지로 전일 대비 일별 변동으로 범주화하여 분석에 투입하였다. 거시경제변수로 미국 달러의 강세를 지수화한 '달러 지수(Dollar_index)'의 일별 데이터를 수집하여 전일 대비 일별 변동으로 범주화하여 분석에 투입하였다.

지금까지는 거시경제와 증시관련 변수였다. 다음으로 코로나19 관련한 투입변수로는 국가통계포털 KOSIS에서 제공하는 코로나19 발생현황의 일별 신규확진자(new_confirm), 누적확진자(agg_confirm), 신규 사망자(new_death), 누적 사망자(agg_death)를 수집하여 분석에 투입하였다. 또한, 일반적인 뉴스나 기사에서 접할 수 있는 코로나19 발생 현황과 더불어 인터넷 포털의 금융 혹은 증권 세션 상에서 제공하는 뉴스도 일정부분 영향이 있을 것으로 판단하였다. 이에 네이버 증권에서 제공되는 뉴스 중 주요 뉴스에 노출되는 기사들을 일자별로 수집하였다. 그 결과 총 19,939개의 원 자료(기사)를 크롤링하였다. 크롤링과 전처리 툴로 python 3.9의 requests module, json, excel을 활용하였다.

해당 기사의 제목과 부제(혹은 첫 단락의 일부)의 내용으로부터 코로나19와 관련이 있을 것으로 판단되는 키워드를 선정하여 해당 키워드의 노출 빈도를 계수하였다. 그리하여 '코로나(Corona)', '제약(Phama)', '바이오(Bio)', '백



<그림 3> 시간 반영 키워드의 도출 과정

신(Vaccine)', '바이러스(Virus)', '확진(Confirm)', '누적(Accumulate)', '사망(Death)', '신규(New)' 등을 키워드 변수로 추출하였다. 코로나 키워드 변수는 초기 확산지역이 중국 우한이기 때문에, 바이러스에 대한 정식 명칭이 정해지지 않았을 때부터 일정시점까지 우한 폐렴과 코로나19가 혼재되어 사용되었음을 착안하여, 이를 동의어로 간주하고 함께 계수하였다. 이러한 변수들은 주식시장의 개·폐장을 반영하지 않은 단순 키워드이기 때문에, 다음 <그림 3>과 같이 사후처리를 하였으며(김유신 등, 2012; 안성원, 조성배, 2010), 단순 키워드들의 일일 기초 통계량은 <표 2>에 정리하였다.

본 연구에서는 단순 빈도만을 고려한 키워드 대신, 당일 종가지수의 변동을 예측하기에 적합한 형태로 변환하였다. <그림 3>과 같이 어제의 장외 시간대에 뉴스 상 노출된 키워드가 t+1기인 다음날(당일)에 반영되도록 처리하였다. 그러므로 단순 키워드 대신에 '어제 장외 키워드', '당일 장전 키워드', '당일 장중 키워드'로 구체화하여 분석에 투입하였다.

본 절을 요약하자면 시계열, 거시경제지표, 국내·외 증시, 코로나19 발생 현황, 증권 뉴스 기사의 코로나19 관련 키워드를 고려하여 변수 추출을 하였고, 분석에 적합하게 정제·파생·범주화·정규화하여 독립변수로 활용하였다.

Ⅳ. 분석 결과

4.1 예측모형의 성능 평가

분석자료의 기간을 3개년(실험1: 코로나19 발생 전·후 포함)과 1개년(실험2: 코로나19 발생이후)으로 구분하였으며, 목표변수(코스피, 코스피200 헬스케어)에 따라 데이터셋을 달리하여 총 4개의 데이터셋으로부터 데이터 웨어하우스를 구성하였다. 분석에는 C4.5 기반의 의사결정나무 알고리즘을 사용하였고, 분석 툴로는 Weka 3.8 버전을 활용하였다. 분석 시에 학습데이터와 검증데이터의 분할 비율을 7:3으로 설정하였다. 이 때 변수의 중요도는 이득비(gain ratio)로 산정하였으며, 상위 10위권 내에 랭크된 변수들을 실험 별로 정리하면 <표 3>과

<표 3> 실험 별 투입변수의 변수 중요도 랭크 상위 10위

구분	실험1 (n=738)		실험2 (n=248)	
	코스피	코스피200 헬스케어	코스피	코스피200 헬스케어
1위	코스닥	코스닥150 헬스케어	코스닥	코스닥150 헬스케어
2위	코스닥150 헬스케어	코스닥	기준금리 변경 여부	장전 확진 키워드
3위	해당 일자	장전 확진 키워드	해당 일자	해당 일자
4위	어제 장외 제약 키워드	장중 사망 키워드	장중 바이러스 키워드	코스닥
5위	코스피200 헬스케어	장전 누적 키워드	코스닥150 헬스케어	코스피
6위	장중 바이러스 키워드	해당 일자	코스피200 헬스케어	장중 코로나 키워드
7위	장중 백신 키워드	코스피	장중 코로나 키워드	해당 년월일
8위	한국은행 기준금리	해당 년월일	해당 년월일	누적 확진자 수
9위	해당 년월일	장중 코로나 키워드	장전 바이러스 키워드	누적 사망자 수
10위	S&P500	S&P500	미국 증시 개장 여부	신규 사망자 수

같다.

본 연구와 같이 목표변수의 데이터 유형이 범주형(명목형)일 때, 모형의 성능 평가 지표로 예측률(prediction rate) 혹은 적중률(hit ratio), 정분류율(true positive rate), 정밀도(precision), 재현율(recall) 등이 활용되며, 이는 혼동 행렬(confusion matrix)을 통해 계산할 수 있다(Burez, 2009; Longade et al., 2013; Song and Lu, 2015). 본 연구에서는 전체 모형의 예측률과 정분류율을 중심으로 예측모형의 성능을 비교・정리하였으며, 분석의 결과는 <표 4>와 같다.

성능 평가 및 비교에 대한 해석은 다음과 같다. 첫째, 개발된 예측모형 중 가장 높은 예측률은 실험2의 코스피200 헬스케어 예측모형에서 나타났다(77.02%). 둘째, 코로나19 상황을 고려하고 이에 관련된 변수를 투입했기 때문에, 실제로 코로나19가 확산되고 창궐된 기간에 해당하는 실험2의 예측률이 실험1에 비해 전반적으로 높다. 셋째, 목표변수의 하위집단에 대한 정분류율을 살펴보면 코스피 예측모형의 경우 실험1과 실험2 모두에서 전일 대비 초과(Up)에 대해 분류가 더 잘 나타났다. 즉, 이러한 변수의 구성을 통해서 코스피 일별 변동을 예측하였을 경우, 상승에 대해서는 예측을 잘하지만 하락에 대해서는 예측을 상대적으로 잘하지 못한다고 해석할 수 있다. 이러한 결과는 코스피를 설명할 수 있는 변수들을 전부 고려하지 않은 점과 코스피200 헬스케어에 비해 이항 집단의 불균형적 분포로 인한 데이터 불균형 문제가 어느 정도 있을 것이라고 추정된다. 넷째, 코스피200 헬스케어의 개별 정분류율은 실험1에 비해 실험2가 모두 높다. 즉, 연구 자료의 범위를 코로나19 발생기간(1개년)으로 한정하여 분석하였을 때 전체 모형의 예측률과 개별 정분류율이 개선되었다. 다섯째, 정밀도와 재현율은 모든 모형에서 0.70 이상의 수준으로 나타났다.

성능 평가 지표와 해석을 종합할 때, 코로나19의 영향을 파악함에 있어 전체 자료의 기간을 3개년으로 했을 때(실험1)보다, 2020년 1개년에 대한 예측모형 성능이 더 우수하다고 볼 수 있다. 기간을 1개년으로 한정하였을 때, 동일한 변수들을 구성하여 분석했음에도 불구하고, 코스피 예측모형보다 코스피200 헬스케어 예측모형의 상승폭이 더 크다는 점도 파악할 수 있다.

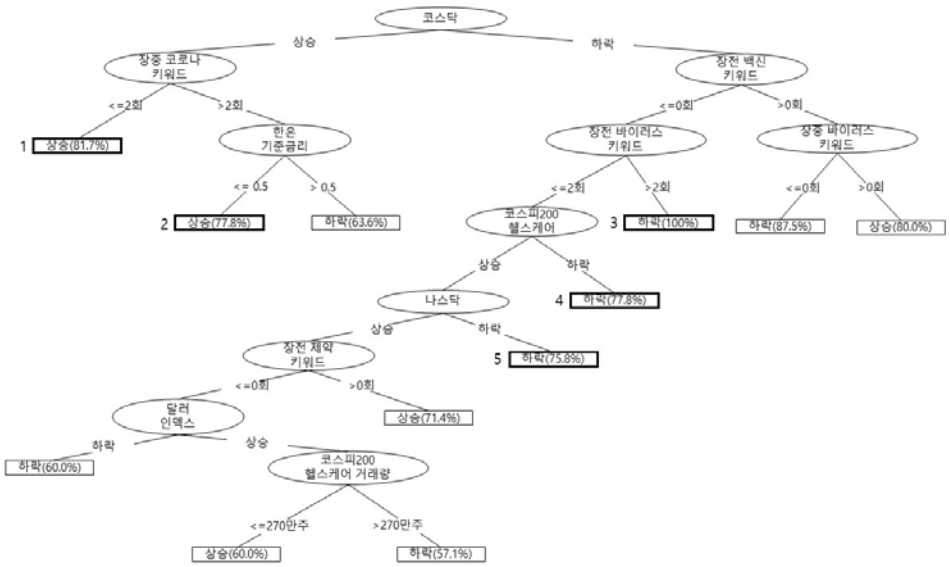
<표 4> 실험 별 예측모형의 정보와 성능 평가

구 분		Test 1 (n=738)		Test 2 (n=248)	
		KOSPI	KSP200 Healthcare	KOSPI	KSP200 Healthcare
독립변수		48개	48개	48개	48개
분할비율		7:3	7:3	7:3	7:3
규칙 도출		12개	14개	10개	12개
예측률		71.04%	70.13%	72.97%	77.02%
정분류율	Up	0.844	0.669	0.848	0.771
	Down	0.545	0.738	0.536	0.769
정밀도		0.716	0.706	0.724	0.771
재현율		0.710	0.701	0.730	0.770

4.2 전체 기간 자료의 실험 결과

4.2.1 실험1의 코스피 의사결정나무와 규칙 도출

첫 번째 실험은 코로나19가 발생하기 전의 2년치 자료와 코로나19가 발생한 후의 자료를 모두 합한 3년간의 자료를 대상으로 코스피와 코스피200 헬스케어 일일 변동을 목표변수로 선정하였다. 분석 자료 전체 기간의 코스피 의사결정나무는 <그림 4>와 같다. 분석 결과, 실험1의 코스피 의사결정나무는 최상단 뿌리노드에 코스닥 일별 변동이 분지 기준으로 나타났다. 코스닥이 전일 대비 상승했을 때에 속하는 규칙이 3개, 유지되거나 하락했을 때에 속하는



<그림 4> 실험1 코스피 의사결정나무

<표 5> 실험1 코스피 일별 변동 규칙

규칙	설명	목표변수	예측 수	적중률 (%)
1	코스닥 일별 변동=Up & 장중 코로나 키워드<=0.2	Up	393	81.6
2	코스닥 일별 변동=Up & 장중 코로나 키워드>0.2 & 한국은행 기준금리<=0.5	Up	9	77.7
3	코스닥 일별 변동=Down & 장전 백신 키워드<=0 & 장전 바이러스 키워드>0.143	Down	28	100
4	코스닥 일별 변동=Down & 장전 백신 키워드<=0 & 장전 바이러스 키워드 <=0.143 & 코스피200 헬스케어 일별 변동=Down	Down	212	77.8
5	코스닥 일별 변동=Down & 장전 백신 키워드<=0 & 장전 바이러스 키워드 <=0.143 & 코스피200 헬스케어 일별 변동=Up & 나스닥 일별 변동=Down	Down	33	75.7

규칙이 9개가 도출되었다. 총 12개의 규칙 중에서 5개의 규칙을 선별하여 정리하면 <표 5>와 같다.

정리된 규칙들 중 1번과 4번 규칙은 규칙 내에 포함된 인스턴스가 많고, 적중률이 77% 이상임으로 한국 증시 3개년(2018~2020년)의 패턴을 잘 반영한 것으로 판단된다. 1번 규칙은 한국 증시를 나타내는 또 다른 지표인 코스닥이 전일 대비 상승했을 때, 장전에 코로나19 키워드가 포함된 뉴스 기사가 2회 이하로 노출된다면 높은 확률로 코스피가 전일 대비 상승했다는 패턴을 의미한다. 이 때 2회라는 해석은 모형에 나타난 정규화된 값을 참고하여, 실제 데이터를 역추적하여 도출된 것이다. 구체적으로, 장전 코로나 키워드 변수의 분지 기준값으로 0.2가 나타났기 때문에, 해당 값을 실제 데이터로 역추적하였을 때 2회임을 확인하였다. 하위 단락에서도 동일하게 정규화된 값이 분지 기준값으로 등장했을 경우, 역추적을 하여 실제 값을 확인하였다.

4번 규칙은 코스닥이 전일 대비 하락세 였고, 장전에 백신 키워드가 포함된 뉴스 기사가 0회 초과로 노출, 장전에 바이러스 키워드가 포함된 뉴스 기사가 2회 이하로 노출, 코스피200 헬스케어가 전일 대비 하락세라면 높은 확률로 코스피가 전일 대비 하락했다는 패턴을 의미한다. 이는 3번 규칙과 매우 유사한 패턴인데, 백신에 대한 기대감이 뉴스 기사로 장전에 투자자들에게 전달되지 않았을 때 코스피 종가는 유지되었거나, 하락했던 것으로 보인다. 오히려 장전에 바이러스에 대한 뉴스 기사가 투자자들에게 전달됨으로, 예측수는 적지만 100%의 적중률을 보인 3번 규칙이 나타났다.

이 밖에도 코로나19 팬데믹 공포로 경기침체를 우려한 당국이 기준금리를 2회에 걸쳐 인하시켰던 점을 상기할 때(한국은행, 2021), 2번 규칙의 분지 기준으로 기준금리가 0.5라는 값이 나타난 부분은 2020년 5월 28일 전·후 상황에 대한 시간적 추론이 가능한 부분이라 할 수 있다[2]. 2번 규칙으로만 한정했을 때, 코스닥이 전일 대비 상승세였고, 장중에 코로나 키워드가 2회 이상 노출된 2020년도 5월 28일 이후의 어느 시점들이 내포된 규칙으로 판단된다. 이것은 김용재와 이상춘(2017)의 연구에서 밝힌 거시경제지표인 총통화, 이자율 등이 주가에 영향을 미친다는 연구와 다소 일치한다. 2020년 한 해 동안 코로나의 위기 속에 경제는 계속해서 위축된 반면, 금리 인하와 통화량의 증가로 인하여 개인 투자자들의 유동성 자금이 주식시장으로 몰리면서 주가는 최고가를 연이어 기록하였다.

끝으로, 미국의 증시 지표 중 하나인 나스닥 일별 변동이 5번 규칙에 포함되었다는 점도 눈여겨 볼 만하다. 이는 미국의 증시가 우리나라의 증시에 영향을 미친다는 김정희와 장병기(2010)의 연구와 일치한다.

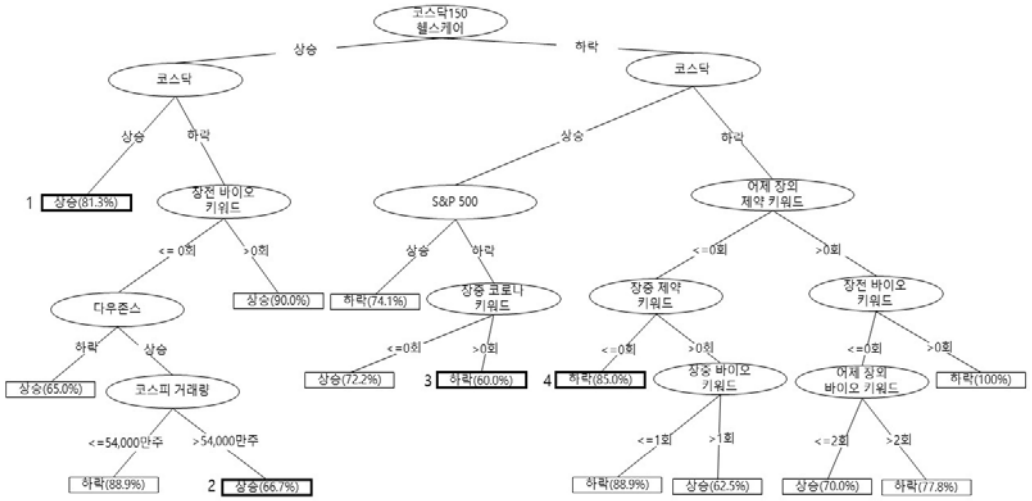
4.2.2 실험1의 코스피200 헬스케어 의사결정나무와 규칙 도출

코스피가 한국 주식시장을 대표하는 지수라면, 코스피200 헬스케어는 굴지의 국내 제약·바이오 기업들이 포함되어 있는 지수라 할 수

2) 코로나19 팬데믹 이후 한국은행 기준금리 변화는 1.25에서 2020년 3월 17일 0.75로 1차 인하, 2020년 5월 28일 0.50으로 2차 인하함.

있다. 본 연구에서는 코스피200 헬스케어가 대표성을 띈다고 판단하고 목표변수로 선정하였다. <그림 5>는 분석 자료 전체 기간의 코스피200 헬스케어 의사결정나무이다.

실험1의 코스피200 헬스케어 의사결정나무는 최상단 뿌리노드에 코스닥150 헬스케어 일별 변동이 분지 기준으로 나타났다. 코스닥150 헬스케어가 전일 대비 상승했을 때에 속하는 규칙이 5개, 유지되거나 하락했을 때에 속하는 규칙이 9개가 도출되었다. 총 14개의 규칙 중에서 4개의 규칙을 선별하여 <표 6>과 같이 정리하였다.



<그림 5> 실험1 코스피200 헬스케어 의사결정나무

<표 6> 실험1 코스피200 헬스케어 일별 변동 규칙

규칙	설명	목표변수	예측수	적중률 (%)
1	코스닥150 헬스케어 일별 변동=Up & 코스닥 일별 변동=Up	Up	331	81.2
2	코스닥150 헬스케어 일별 변동=Up & 코스닥 일별 변동=Down & 장중 바이오 키워드<=0 & 다우존스 일별 변동=Up & 코스피일일 거래량>0.19	Up	9	66.7
3	코스닥150 일별 변동=Down & 코스닥 일별 변동=Up & S&P500 일별 변동=하락 & 장중 코로나 키워드>0	Down	10	60.0
4	코스닥150 일별 변동=Down & 코스닥 일별 변동=Down & 어제 장외 제약 키워드<=0 & 장중 제약 키워드<=0	Down	234	85.0

정리된 규칙들 중에서 1번과 4번 규칙은 규칙 내에 포함된 인스턴스가 많고, 적중률이 81% 이상임으로 유의미한 규칙이라 판단된다. 국내 제약·바이오 분야에 대한 또 다른 증시 지표인 코스닥150 헬스케어의 작동과 더불어 코스닥의 작동에 의해 규칙이 달라짐을 알 수 있다. 코스닥 시장에서 코스닥150 헬스케어에 속하는 기업들이 2020년 매출액 순위가 크게

상승한 점과 시가총액 상위를 석권한 것을 상기해볼 때[3][4], 3개년에 대한 예측모형임에도 불구하고 이들 변수들 간의 매우 큰 상관관계가 존재하는 것으로 판단된다.

1번 규칙은 코스닥150 헬스케어가 상승세이며, 코스닥이 상승세일 때 코스피200 헬스케어가 높은 확률로 상승했음을 반영한 규칙이다. 4번 규칙은 코스닥150 헬스케어와 코스닥이 하락세일 때, 제약 키워드가 뉴스에 노출되지 않으면 높은 확률로 코스피200 헬스케어가 하락했음을 반영한 규칙이다. 이는 고승의(2018)의 연구 결과를 다소 뒷받침한다. 그의 연구에서 제약·바이오 관련 성장주의 경우 호재가 있으면 일시적으로 가치주보다 높은 성장을 기록했다고 했다. 이러한 현상이 코로나19로 인해 발생한 팬데믹 현상에서도 나타났다고 볼 수 있다.

제약 키워드 변수의 경우, 시간별로 어제 장외 변수와 장중 변수가 조합됨으로 규칙이 도출되었다. 이는 시간을 고려하지 않고 단순 키워드 변수를 구성했을 때와 달리, 시간을 고려함으로 해석적 측면의 풍부함을 더해줄 수 있는 연구방법론의 장점이 나타난 부분이라 할 수 있다. 이 규칙이 시사하는 바는 성노윤과 남기환(2018)의 연구에서 밝힌 바와 같이 온라인 뉴스가 주가 예측에 있어 주요하며, 개장된 주식시장 내에서의 시간적 요소(장전·장중·장마감 후)를 고려해야 한다는 주장을 지지한다.

이 밖에 2번, 3번 규칙은 다소 적중률이 떨어지는 규칙들이지만, 미 증시를 대표하는 S&P 500과 다우존스 지수, 코스피 일일 거래량이 규칙 내의 독립변수 조합으로 나타났다. 또한 코스피 일일 거래량의 분지 기준점인 0.19라는 정규화된 값을 역추적하면 0.19에 해당되는 실제 값의 구간을 발견할 수 있고, 평균 5억4천3백만 주로 실제 값을 특정할 수 있다[5].

4.3 팬데믹 기간 자료의 실험 결과

4.3.1 실험2의 코스피 의사결정나무와 규칙

한국 주식시장 및 헬스케어 산업과 코로나19의 영향 관계를 파악하기 위해서, 전체 자료의 기간이 아닌 코로나19가 본격적으로 창궐했던 2020년 1개년으로 범위를 한정하여 실험2를 진행하였다. 앞서 언급한 바와 같이 한국 주식시장을 대표하는 지수인 코스피와 국내 제약·바이오 분야를 대변할 수 있는 코스피200 헬스케어를 각각 목표변수로 선정하였다.

먼저, 분석자료 중 부분적 기간의 코스피 의사결정나무는 <그림 6>과 같다. 실험2의 코스피 의사결정나무도 최상단 뿌리노드에 코스닥 일별 변동이 분지 기준으로 나타났다. 코스닥이 전일 대비 상승했을 때에 속하는 규칙이 3개, 유지되거나 하락했을 때에 속하는 규칙이 7개가 도출되었다. 총 10개의 규칙 중에서 6개의

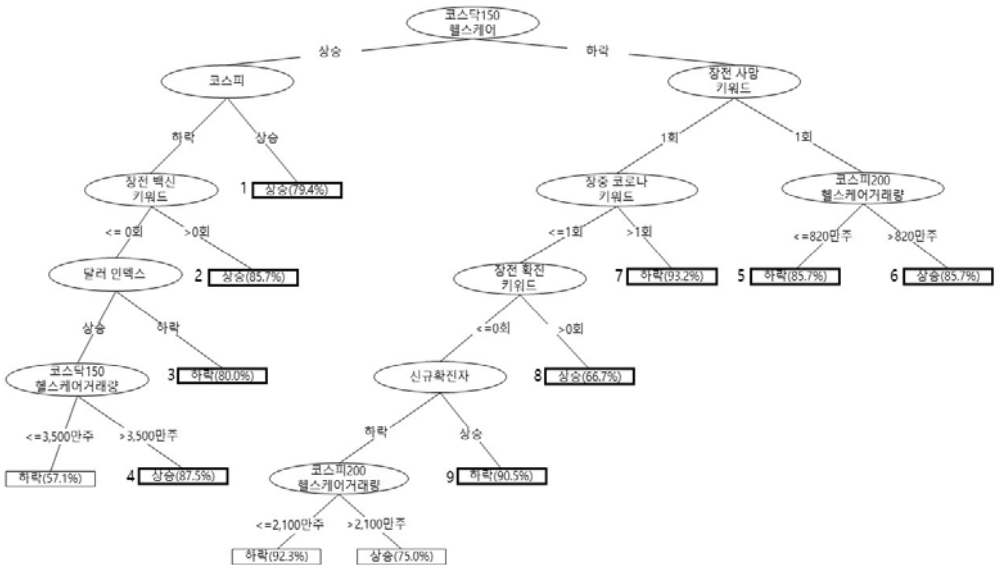
3) 의약뉴스 “코스닥 시장 제약·바이오업체 매출액 순위 약진,” http://www.newsmp.com/news/articleView.html?idxno=213962

4) 팜스탁 “제약바이오, 코스닥 시총 상위 ‘싹슬이’,” http://www.pharmstock.co.kr/news/articleView.html?idxno=27839

5) KOSPI 일일 거래량의 z값이 0.19인 실제 값의 구간은 최소 551,703,792 ~ 최대 552,586,618 주로 해당 구간에 포함되는 관측 수는 13개임. 이를 평균한 값은 543,160,524 주로 <표 2>의 KOSPI 일일 거래량의 전체 산술평균과 근접한 부분이라는 점에서 오른쪽 꼬리 분포임을 추론할 수 있음.

규칙을 선별하여 <표 7>과 같이 정리하였다.

전반적으로 실험1의 코스피 예측모형과 실험2의 코스피 예측모형은 의사결정나무의 모양과 도출된 규칙, 분지 기준 등이 유사함을 살펴볼 수 있다. 그러나 2020년 한 해는 코로나19가 창궐한 시기인 만큼, 코로나19 발생 현황에서 사용되는 단어인 '확진'이 포함된 키워드 변수가 분지 기준으로 새롭게 나타났다. 특히, 실험1의 1번 규칙과 실험 2의 1번 규칙의 경우, 독립변수의 조합이 동일한데 이에 관련해서 두 가지 합리적인 추론이 가능하다.



<그림 6> 실험2 코스피 의사결정나무

<표 7> 실험2 코스피 일별 변동 규칙

규칙	설명	목표변수	예측수	적중률 (%)
1	코스닥 일별 변동=Up & 장중 코로나 키워드<=0.2	Up	140	85.7
2	코스닥 일별 변동=Up & 장중 코로나 키워드>0.2 & 한국은행 기준금리<=0.5	Up	9	77.8
3	코스닥 일별 변동=Up & 장중 코로나 키워드>0.2 & 한국은행 기준금리>0.5	Down	11	63.6
4	코스닥 일별 변동=Down & 장중 백신 키워드>0	Up	9	55.6
5	코스닥 일별 변동=Down & 장중 백신 키워드<=0 & 장전 확진 키워드>0	Down	21	95.2
6	코스닥 일별 변동=Down & 장중 백신 키워드<=0 & 장전 확진 키워드<=0 & 장전 백신 키워드<=0 & 장전 바이러스 키워드>0.143	Down	15	100.0

첫째, 20년도 한 해 동안만 언급된 코로나라는 키워드가 끼치는 영향력이 자료의 기간을 3개년으로 확대하였음에도 불구하고 매우 컸음을 알 수 있다. 즉, 코로나19라는 바이러스와 질병이 없었던 기간이 분석 자료 내에서만 약 490여일이 포함되어 있음에도, 장중에 코로나 키워드가 기사화되었는지 여부에 따라 규칙이 달라진다. 즉, 실험1과 2의 1번 규칙은 코스닥 상승세와 함께 장중에 코로나 키워드가 2회 이하로 노출되면 코스피가 전일 대비 상승했음을 의미하며, 높은 적중률과 많은 양의 예측수를 포함하고 있는 규칙이기 때문에 매우 단순하지만 지난 3년과 1년간의 한국 증시와 코로나19의 영향을 잘 반영한 패턴이라 평가할 수 있다.

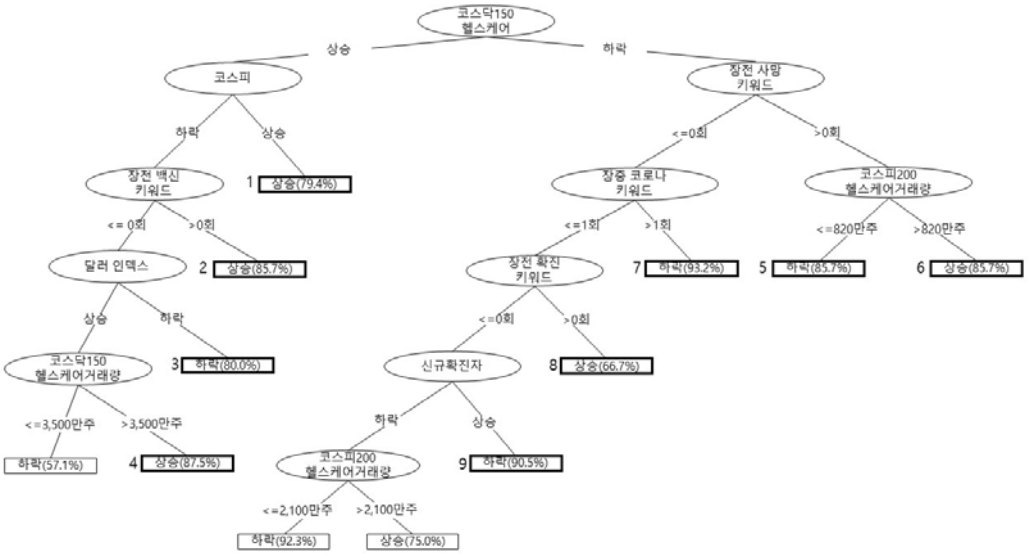
둘째, 동일한 독립변수의 조합을 가진 규칙이라도 자료의 기간에 따라 적중률 차이가 약 5% 가량 차이가 나타났다. 이는 마찬가지로 코로나19의 창궐이 본격적으로 시작된 부분적 기간(2020년)만을 고려하였을 때 높은 예측 정확도를 보인 것은 코로나19 관련 변수의 효과가 얼마나 큰지를 알 수 있다.

<표 7>의 결과를 살펴보면, 실험1에 비하여 실험2는 규칙 내에 포함되는 예측수가 감소하였음을 알 수 있다, 이는 1개년 데이터만을 활용한 것이 반영된 결과이다(n=248). 1번은 규칙의 적중률은 85% 이상이며, 절반 이상의 인스턴스가 리프 노드에 포함된다. 5번 규칙은 적중률이 95% 이상으로 매우 정확한 예측 규칙으로 간주될 수 있다. 또한 1번 규칙 내에 포함되지 않은 남은 인스턴스의 수가 108개임을 감안할 때, 예측수가 21개로 고려할만한 규칙으로 판단된다. 5번 규칙의 경우 주식시장에서 뉴스나 WOM(Words Of Mouth)의 효과가 주가에 영향을 미치는 중요한 변수임을 시사한다. 이는 주가 분석에 있어 증권 뉴스에 대한 텍스트 마이닝 활용이 효과가 있다는 김유신 등(2012), 홍태호 등(2017)의 연구를 입증한다.

4.3.2 실험2의 코스피200 헬스케어 의사결정나무와 규칙

2020년 한 해 동안의 코스피200 헬스케어 분석자료를 활용한 부분적 기간의 코스피(코스피200 헬스케어) 의사결정나무는 <그림 7>과 같다. 실험2의 코스피200 헬스케어 의사결정나무도 최상단 뿌리노드에 코스닥150 헬스케어 일별 변동이 분지 기준으로 나타났다. 코스닥150 헬스케어가 전일 대비 상승했을 때에 속하는 규칙이 5개, 유지되거나 하락했을 때에 속하는 규칙이 7개가 도출되었다. 총 12개의 규칙 중에서 9개의 규칙을 선별하여 <표 8>과 같이 정리하였다.

실험1의 코스피200 헬스케어 예측모형과 실험2의 코스피200 헬스케어 예측모형을 비교해보면, 분지 기준으로 나타난 키워드 변수들의 변화가 뚜렷함을 알 수 있다. 전체 기간(3개년)인 실험1의 예측모형은 ‘제약’, ‘바이오’, ‘코로나’ 키워드가 분지 기준으로 나타났다. 반면 부분적 기간(1개년)인 실험2의 예측모형은 제약과 바이오가 나타나지 않고, ‘백신’, ‘코로나’, ‘확진’, ‘사망’ 키워드가 분지 기준으로 나타났다. 특히, ‘확진’, ‘사망’이라는 키워드는 증권 뉴스 상에 기사 뿐만 아니라 여러 매체를 통해 전파되는 코로나19 발생현황에서 사용되는 단어이기 때문에 의미가 있다고 판단된다. 또한 코로나19 발생 현황을 통해 추출한 변수인 ‘신규확진자’가 모든 예측모형 중에서 유일하게



<그림 7> 실험2 코스피200 헬스케어 의사결정나무

분지의 기준으로 나타났다는 점은 눈 여겨 볼 만한 부분이다. 이것은 2020년 한 해 동안 코로나 확진이나 사망과 같은 충격적인 키워드가 얼마나 주식시장에 큰 파장을 일으키는가를 단적으로 보여준다.

<표 8>의 결과를 살펴보면, 규칙들 중 8번을 제외한 규칙들의 적중률이 약 80%를 상회하며, 7번 규칙과 9번 규칙은 90% 이상임으로 소표본 대비 포함된 예측수가 적지 않아서 매우 유의미한 규칙이라 판단된다. 1번 규칙은 코스닥150 헬스케어의 상승세와 코스피200 헬스케어가 속한 코스피가 상승세일 때, 높은 확률로 코스피200 헬스케어가 전일 대비 상승을 했음을 보여주는 패턴이다. 2번 규칙은 코스닥150 헬스케어가 상승세, 코스피가 전일 대비 하락세 중에 장전에 백신 키워드가 1회 이상 증권 뉴스상에 노출이 되면, 높은 확률로 코스피200 헬스케어가 상승을 했음을 보여주는 규칙이다.

3-4번 규칙은 주가를 예측하거나 시장 환경을 반영하기 위해 활용되는 거시경제변수 중 하나로 환율을 고려한 연구들(김정희, 장병기, 2010; 김주일, 2013; 고강석, 2018)을 지지하는 규칙으로, 미국 달러의 강세를 나타내는 달러지수가 독립변수의 조합에 포함된다. 환율의 통계적 유의성이 연구마다 다르다는 점과 정(+)과 부(-)의 영향도 다르다는 점을 참고할 때, 데이터에 입각한 추론의 어려움으로 향후 연구를 통해 보완하고자 한다.

5-6번 규칙은 코스닥150 헬스케어의 하락세, 장전 사망자 키워드가 증권 뉴스에 1회 이상 노출이라는 조합에서 코스피200 헬스케어 일일 거래량에 따라 코스피200 헬스케어 일일 변동이 다르게 나타난다. 이때, 기준값인 0.05는 실제 값들의 구간 평균인 8백2십만 주 정도로 추정이 된다. 즉, 코스피200 헬스케어의 일일 거래량이 8백2십만 주보다 초과로 거래되면, 코

<표 8> 실험2 코스피200 헬스케어 일별 변동 규칙

규칙	설명	목표변수	예측수	적중률 (%)
1	코스닥150 헬스케어 일별 변동=Up & 코스피일별 변동=Up	Up	107	79.4
2	코스닥150 헬스케어 일별 변동=Up & 코스피일별 변동=Down & 장전 백신 키워드>0	Up	7	85.7
3	코스닥150 헬스케어 일별 변동=Up & 코스피일별 변동=Down & 장전 백신 키워드<=0 & 달러지수 일별 변동=Down	Down	10	80.0
4	코스닥150 헬스케어 일별 변동=Up & 코스피일별 변동=Down & 장전 백신 키워드<=0 & 달러지수 일별 변동=Up & 코스닥150 헬스케어 일일 거래량>0.26	Up	8	87.5
5	코스닥150 헬스케어 일별 변동=Down & 장전 사망 키워드>0 & 코스피200 헬스케어 일일 거래량<=0.05	Down	7	85.7
6	코스닥150 헬스케어 일별 변동=Down & 장전 사망 키워드>0 & 코스피200 헬스케어 일일 거래량>0.05	Up	7	85.7
7	코스닥150 헬스케어 일별 변동=Down & 장전 사망 키워드<=0 & 장중 코로나 키워드>0.1	Down	44	93.2
8	코스닥150 헬스케어 일별 변동=Down & 장전 사망 키워드<=0 & 장중 코로나 키워드<=0.1 & 장전 확진 키워드>0	Up	9	66.7
9	코스닥150 헬스케어 일별 변동=Down & 장전 사망 키워드<=0 & 장중 코로나 키워드<=0.1 & 장전 확진 키워드<=0 & 신규 확진자 일별 변동=Up	Down	21	90.5

스피200 헬스케어의 종가지수가 전일 대비 상승할 확률이 높다. 반대로 거래량이 8백2십만 주 이하로 거래되면 종가지수가 전일과 동일하게 유지되거나 하락할 확률이 높다. 두 규칙 모두 예측수가 7개로 다소 일반화하기엔 어려운 규칙이긴 하지만, 거래량이라는 변동성에 의해 분지가 된다는 점과 규칙이 달라진다는 점에서 의미가 있다. 코로나19 상황, 특히 사망자 관련 뉴스가 장전에 노출되었음에도 불구하고 특정 거래량 이상이면 코스피200 헬스케어의 종가지수가 상승했다는 점은 높은 전파력과 사망으로까지 이르게 하는 질병을 치료할 수 있는 백신의 잠정적 개발사들에 대한 '강세장 속 투자자들의 기대심리'가 반영된 패턴으로 보인다 (김명진 등, 2020).

7번과 9번 규칙은 장세가 전일 대비 좋지 않고, 코로나19를 반영한 변수들의 어떠한 작동에 의해서 매우 높은 확률로 코스피200 헬스케어가 전일 대비 하락한다는 패턴을 보여준다. 사망 키워드가 장전 뉴스에 노출되지 않은 상태에서 장중 뉴스에 코로나가 1회를 초과해서 노출되면, 코스피200 헬스케어가 전일 대비 하락하는 패턴이 7번 규칙에 해당된다. 장전 사망 키워드, 장중 코로나 키워드, 장전 확진 키워드가 희박하게 증권 뉴스의 기사에 노출되더라도, 공표되는 코로나19 발생자 현황의 신규확진자 수가 전일 대비 상승하면, 코스피200 헬스케어가 전일 대비 하락하는 패턴이 9번 규칙에 해당된다. 즉, 7번과 9번은 코로나19를 반영하기 위한 변수로 고려했던 변수들의 세부적 조합에 의해 규칙이 도출된 것이다. 즉, 코로나19 관련 키워드가 증권 뉴스에 노출되거나, 코로나19

신규확진자 수가 전날에 비해 증가했음이 어떠한 매체를 통해 발표됨으로 인해, 매우 높은 확률(90% 이상)로 코스피200 헬스케어가 유지 혹은 하락했음을 보여주는 패턴이라 할 수 있다. 이것은 팬데믹 관련 키워드들이 확실히 주식시장에 영향을 미친다는 것을 보여준다.

Ⅴ. 결 론

본 연구는 코로나19가 국내 주식시장과 제약·바이오 산업에 대한 영향을 규명하기 위하여 시도된 연구로 자료 수집과 패턴 분석을 위해 텍스트 마이닝과 데이터 마이닝 기법을 활용하였다. 분석을 위해 이자율, 기준금리 등과 같은 다양한 산업경제 거시변수는 물론이며 미국의 증권시장 지수도 동시에 활용하였다. 특히, 본 연구에서는 정부에서 매일 발표되는 코로나19 관련 통계치와 증권시장의 주가에 영향을 미칠 수 있는 코로나19 관련 키워드들을 웹으로부터 크롤링하였으며 이를 계량화하여 분석에 사용한 것이 타 연구와는 상당히 차별된다.

기존의 통계적 방법론 중 주가 예측에 많이 활용되는 선형회귀모형이 인과 관계에 기반한 분석 방법이라면, 본 연구에서 사용한 데이터 마이닝은 상관관계에 기반한 분석 방법이라 할 수 있다. 각 독립변수가 종속변수에 미치는 영향의 구체적인 크기에 대해서는 본 연구방법론을 통해 파악하기는 어렵지만, '독립변수들의 조합'에 의해 패턴을 발견하고, 예측 가지수와 적중률을 통해 발견된 패턴이 얼마나 유용한지를 설명하였다. 설정한 연구 문제에 대한 해답을 제시하기 위하여 4개의 실험세트가 만들어졌으며, 각 실험마다 각각 하나의 의사결정나무 기반의 예측모형과 각 모형별로 10~14개의 의사결정규칙을 도출했다. 이를 활용하여 '전체 기간의 국내 증시와 헬스케어 분야 증시의 패턴'과 '코로나19 기간의 국내 증시와 헬스케어 분야 증시의 패턴'을 시각화하고 비교·정리하였다.

연구 결과를 정리하면 다음과 같다. 첫째, 기술통계량에 대한 해석을 통해 2020년 한 해 증시가 매우 급변하였다는 점을 파악하였다. 연구 자료의 전체 범위 내에서 코스피 및 코스피200 헬스케어의 최저치와 최고치를 모두 2020년에 기록하였다. 즉, 발병과 창궐 시기인 2020년은 물론이고 최근 3년간으로 범위를 확장하더라도, 코로나19가 한국 증시에 매우 큰 충격을 주었다고 할 수 있다.

둘째, 실험1을 통해, 최근 3년간의 증시에 대한 전반적인 패턴을 도식과 규칙을 통해 정리하였다. 또한 실험2를 통해, 코로나19 대유행과 공포, 백신에 대한 기대 심리 등으로 급변하는 시기를 맞이했던 2020년의 증시에 대한 전반적인 패턴을 도식과 규칙을 통해 정리하였다.

셋째, 분석에 투입된 변수들이 목표변수와 실험에 따라 대동소이하게 나무모형 상에 분지 기준으로 나타났으며, 도식과 규칙을 정리하고 이를 충실히 해석하였다. 특히, 증시 개·폐장 시간을 고려한 키워드 변수를 활용하여 투자자들이 직접적으로 뉴스를 접할 수 있는 시간대를 반영하였기 때문에 보다 논리적이고 풍부한 해석과 추론을 시도할 수 있었다.

넷째, 정부에서 매일 발표하는 코로나19 확진자와 사망자 통계는 시시각각으로 변화하는

팬데믹의 공포를 안겨주었고, 이것은 곧바로 주가에 반영되었다. 즉, 코로나19에 대한 정부의 공식적인 통계는 매우 신뢰성 있는 데이터로 작용하여 증권시장에 곧바로 반영되었다.

다섯째, 증권시장에 떠도는 코로나19 치료제나 백신에 관한 뉴스들은 제약・바이오 기업들에게 큰 호재가 되어 전체적인 경기 침체에도 불구하고 이들 기업에게는 큰 기회요인으로 작용하였다. 이것은 일반적인 상황보다 팬데믹 현상과 같은 이상현상이 발생할 경우 뉴스 정보와 같은 텍스트 정보를 가공하여 예측하는 텍스트 마이닝 기법이 매우 유용함을 시사한다.

여섯째, 이자율이나 기준금리와 같은 거시경제변수는 항상 자산시장에 큰 영향을 미치는 변수이다. 본 연구에서는 코로나19 극복과 경기 활성화를 위해 정부가 취한 기준금리인하 정책은 증시부양수준을 넘어 개미 투자자들의 투자과열로 이어졌다. 이로 인하여 2020년 한 해 동안 우리나라 증권시장은 역사상 유래 없이 가장 변동성이 높은 한 해로 기록되었다. 거시변수를 통한 정부의 시장개입은 항상 처음 예측한 것보다 더 크게 시장에서 반향을 불러일으키는 경우가 많았다. 이러한 현상은 혼란과 혼돈의 시대에 더 크게 작용한다는 것을 본 연구는 실증적으로 보여주고 있다.

마지막으로 우리 경제는 미국과 같은 선진국 경제와 항상 맥을 같이한다. 특히, 코로나19와 같은 팬데믹 현상이 발생할 경우 더욱 잘 나타났다. 2007년-2008년 미국의 금융위기 때 우리나라 증권시장도 큰 폭으로 같이 움직였다. 이번 팬데믹 상황에서도 미국의 증시가 우리나라의 증시에 큰 영향을 미쳤다. 이러한 사실을 본 연구에서도 동일하게 검증하였다.

본 연구의 한계점은 다음과 같다. 첫째, 데이터 마이닝 분석기법만을 활용하였기 때문에 해석과 추론에 있어 일각에선 과대 해석으로 판단할 수도 있다. 향후 연구를 통해 인과적 분석을 병행하여 해당 부분을 보완하고자 한다. 둘째, 코스피는 한국 증시를 모두 포함하는 지표임에도 불구하고 제한적인 변수들만 고려하였기 때문에, 전 업종별 주가지수를 고려하지 않았으며 한정된 거시경제변수들을 사용하였다. 향후, 생산자 물가지수, 소비자 물가지수, 국제유가, 국제 수지, 경기실사지수, 업종별 주가지수 등의 변수들을 추가함으로 좀 더 실세계를 반영한 데이터 분석을 시도하고자 한다.

참고문헌

강현철 외 5명, 빅데이터 분석을 위한 데이터마이닝 방법론, 자유아카데미, 2014.

고강석, “업종별 주가지수와 주가-환율 관계의 안정성,” Journal of Korean Data Analysis Society, 제20권, 제2호, 2018, pp. 815-828.

고승의, “한국 주식시장에서 성장주와 가치주의 초과수익률 비교,” 한국융합학회논문지, 제9권, 제7호, 2018, pp. 213-222.

김덕현, 유동희, 정대율, “의사결정나무 기법을 이용한 노인들의 자살생각 예측모형 및 의사결정 규칙 개발,” 정보시스템연구, 제28권, 제3호, 2019, pp. 249-276.

김명진, 류지혜, 차동호, 심민규, “SNS감성 분석을 이용한 주가 방향성 예측: 네이버 주식토론방 데이터를 이용하여,” 한국

전자거래학회지, 제25권, 제4호, 2020, pp. 61-75.

김민정, 권택호, “코로나19와 한국 제약바이오 기업의 시장성과,” 경영경제연구, 제43권, 제1호, 2021, pp. 165-187.

김용재, 이상춘, “한국의 거시경제 지표가 기업의 주가수익률에 미치는 영향분석: SVAR 활용,” 전문경영인연구, 제20권, 제3호, 2017, pp. 281-303.

김유신, 김남규, 정승렬 , “뉴스와 주가: 빅데이터 감성분석을 통한 지능형 투자결정모형,” 지능정보연구, 제18권, 제2호, 2012, pp. 143-156.

김정희, 장병기, “금융시장 환경변화에 따른 거시경제 및 해외변수들의 주가에 대한 영향력 변화,” 산업경제연구, 제23권, 제4호, 2010, pp. 2105-2134.

김주일, “KOSPI 지수 및 KOSDAQ 지수와 환율과의 상호연관성에 관한 연구,” 한국산업경제학회 2013년도 추계국제학술대회 논문집, 2013, pp. 37-54.

손애린, 신왕수, 이준기, “텍스트 마이닝 기반의 자산관리 핀테크 기업 핵심 요소 분석: 사용자 리뷰를 바탕으로,” 정보시스템연구, 제29권, 제4호, 2020, pp. 137-151.

성노윤, 남기환, “산업군 내 동질성을 고려한 온라인 뉴스 기반 주가 예측,” 지능정보연구, 제24권, 제2호, 2018, pp. 1-19.

안성원, 조성배, “뉴스 텍스트 마이닝과 시계열 분석을 이용한 주가 예측,” 한국정보과학회 2010 한국컴퓨터종합학술대회 논문집, 제37권, 제1호, 2010, pp. 364-369.

유지돈, 이익선, “인터넷 뉴스 콘텐츠 빅데이터를 활용한 종합주가지수 예측,” 경영과학, 제35권, 제4호, 2018, pp. 1-14.

이용대, 박현기, 김수욱, “한국 주식시장에서 주식 수익률과 거래량 간의 동적 관계 분석,” 산업경제연구, 제28권, 제2호, 2015, pp. 739-758.

전새미, 정여진, 이동엽, “개별 기업에 대한 인터넷 검색량과 주가변동성의 관계: 국내 코스닥시장에서의 산업별 실증분석,” 지능정보연구, 제22권, 제2호, 2020, pp. 81-96.

전해정, “코로나-19가 자산시장에 미친 영향에 관한 연구,” 인문사회 21, 제11권, 제5호, 2020, pp. 2059-2070.

정용완, 정승국, “데이터웨어하우스(DW) 어플라이언스 기술 동향 분석,” 한국정보기술학회지, 제10권, 제2호, 2012, pp. 107-111.

홍태호, 김은미, 차은정, “뉴스 감성분석과 SVM을 이용한 다우존스 지수와 S&P500 지수 예측,” 인터넷전자상거래연구, 제17권, 제1호, 2017, pp. 23-36.

KOSIS, “코로나 19 발생현황,” 2020, https://kosis.kr/covid/covid_index.do

WHO, “Coronavirus Disease (COVID-19) Report,” 2020, https://www.who.int/emergencies/diseases/novel-coronavirus-2019

한국거래소, “주식시장-유가시장-유가증권시장 소개-코스피,” 2021a, http://open.krx.co.kr/contents/OPN/01/01010101/OPN

01010101.jsp

한국거래소. “파생상품시장-주가지수상품-코스피200 섹터지수선물,” 2021b, http://open.krx.co.kr/contents/OPN/01/01040203/OPN01040203.jsp

한국은행, “통화정책-통화정책방향-한국은행 기준금리 추이,” 2021, https://www.bok.or.kr/portal/singl/baseRate/list.do?dataSeCd=01&menuNo=200643

Burez, J. and Van den Poel, D., “Handling Class Imbalance in Customer Churn Prediction,” *Expert Systems with Applications*, Vol. 36, No. 3, 2009, April, pp. 4626-4636.

Longade, R., Dongre, S., and Malik, L., “Class Imbalance Problem in Data Mining: Review,” *International Journal of Computer Science and Network*, Vol. 2, Iss. 1, 2013, February, pp. 1-6.

Song, Y. and Lu, Y., “Decision Tree Methods: Applications for Classification and Prediction,” S*hanghai Arch Psychiatry*, Vol. 27, No. 2, 2015, pp. 130-135.

김 덕 현 (Kim, Deok Hyun)

경상대학교 경영정보학과에서 경영학사와 석사를 취득하였다. 현재 경상국립대학교 경영정보학과에서 박사과정을 수료했다. 주요 관심분야는 빅데이터분석, 노인복지 등이다.

유 동 희 (Yoo, Dong Hee)

고려대학교에서 경영학사와 경영학 박사학위를 취득하였다. 현재 경상국립대학교 경영정보학과에서 부교수로 재직하고 있으며, 주요 관심분야 데이터마이닝, 빅데이터분석, 지능형시스템 등이다.

정 대 율 (Jeong, Dae Yul)

부산대학교 경영학사, 석사, 박사학위를 취득하였다. 현재 경상국립대학교 경영정보학과 교수로 재직하고 있으며, 주요 관심분야는 시스템분석 및 설계, 데이터마이닝, 의사결정지원시스템 등이다.

<Abstract>

Using Data Mining Techniques for Analysis of the Impacts of COVID-19 Pandemic on the Domestic Stock Prices: Focusing on Healthcare Industry

Kim, Deok Hyun · Yoo, Dong Hee · Jeong, Dae Yul

Purpose

This paper analyzed the impacts of domestic stock market by a global pandemic such as COVID-19. We investigated how the overall pattern of the stock market changed due to the impact of the COVID-19 pandemic. In particular, we analyzed in depth the pattern of stock price, as well, tried to find what factors affect on stock market index(KOSPI) in the healthcare industry due to the COVID-19 pandemic.

Design/methodology/approach

We built a data warehouse from the databases in various industrial and economic fields to analyze the changes in the KOSPI due to COVID-19, particularly, the changes in the healthcare industry centered on bio-medicine. We collected daily stock price data of the KOSPI centered on the KOSPI-200 about two years before and one year after the outbreak of COVID-19. In addition, we also collected various news related to COVID-19 from the stock market by applying text mining techniques. We designed four experimental data sets to develop decision tree-based prediction models.

Findings

All prediction models from the four data sets showed the significant predictive power with explainable decision tree models. In addition, we derived significant 10 to 14 decision rules for each prediction model. The experimental results showed that the decision rules were enough to explain the domestic healthcare stock market patterns for before and after COVID-19.

Keyword: COVID-19, KOSPI, Data Mining, Text Mining, Decision Tree, Healthcare Industry

* 이 논문은 2021년 6월 13일 접수, 2021년 7월 29일 1차 심사, 2021년 9월 3일 게재 확정되었습니다.