Journal of Digital Convergence
Vol. 17. No. 9, pp. 79-88, 2019

ISSN 1738-1916
https://doi.org/10.14400/JDC.2019.17.9.079

데이터 마이닝 기법을 활용한 항공기 사고 및 준사고로 인한 사망 발생 요인 및 패턴 분석

김정훈[1], 김태운[2], 유동희[3*]
[1]한국항공우주산업 개발본부 책임연구원,
[2]경상대학교 경영정보학과 학부생, [3]경상대학교 경영정보학과 부교수

Analysis of the Factors and Patterns Associated with Death in Aircraft Accidents and Incidents Using Data Mining Techniques

Jeong-Hun Kim[1], Tae-Un Kim[2], Dong-Hee Yoo[3*]
[1]Senior Research Engineer, Dept. of R&D, Korea Aerospace Industries
[2]Undergraduate Student, Dept. of Management Information Systems, Gyeongsang National University
[3]Associate Professor, Dept. of Management Information Systems, Gyeongsang National University

요 약 본 연구에서는 데이터 마이닝 기법을 활용하여 항공기 사고와 준사고로 인한 사망 발생 요인들과 패턴들을 분석하고자 한다. 이를 위해, 항공기 사고와 준사고 데이터를 보유하고 있는 미국연방교통안전위원회(NTSB)와 미국연방항공청(FAA)의 데이터를 사용하였다. 다음으로 의사결정나무 알고리즘을 사용하여 항공기 사고 및 준사고에 따른 사망여부 예측모형들을 구축하였고 이를 토대로 사망 발생에 영향을 주는 주요 요인들과 패턴들을 도출하였다. NTSB 데이터의 경우 항공기가 완파되거나 고기동 또는 고위험 임무를 수행할 때 주로 사망이 발생하는 것을 알 수 있었다. FAA 데이터의 경우 항공기가 일부 파괴된 경우 조종사의 숙련도가 저조하거나 미인가 조종사의 경우 사망이 발생하였으며, 고공낙하점프와 지상운용단계에서 발생되는 다양한 사망관련 패턴들도 발견되었다. 또한 도출된 패턴들을 활용하여 사망 사고 예방을 위한 실용적인 방안들을 제시한 점에서 연구의 의의를 찾을 수 있다.

주제어 : 데이터 마이닝, 의사결정나무, 예측모형, 항공기 사고, 항공기 준사고

Abstract This study analyzes the influential factors and patterns associated with death from aircraft accidents and incidents using data mining techniques. To this end, we used two datasets for aircraft accidents and incidents, one from the National Transportation Safety Board (NTSB) and the other from the Federal Aviation Administration (FAA). We developed our prediction models using the decision tree classifier to predict death from aircraft accidents or aircraft incidents and thereby derive the main cause factors and patterns that can cause death based on these prediction models. In the NTSB data, deaths occurred frequently when the aircraft was destroyed or people were performing dangerous missions or maneuver. In the FAA data, deaths were mainly caused by pilots who were less skilled or less qualified when their aircraft were partially destroyed. Several death-related patterns were also found for parachute jumping and aircraft ascending and descending phases. Using the derived patterns, we proposed helpful strategies to prevent death from the aircraft accidents or incidents.

Key Words : Data Mining, Decision Tree, Prediction Model, Aircraft Accident, Aircraft Incident

*This article is based on a part of the first author's master's thesis from Gyeongsang National University.
*Corresponding Author : Donghee Yoo(dhyoo@gnu.ac.kr)
Received June 26, 2019 Revised July 26, 2019
Accepted September 20, 2019 Published September 28, 2019

1. 서론

항공기는 3차원 입체교통수단으로 현존하는 가장 빠른 운송매개체로, 여객, 운송, 소방, 의료 및 국방 등 다양한 분야에서 활용되고 있다. 이와 같이 항공 운송수단의 활용성이 증대되는 만큼 항공관련 기술 개발 또한 활발하게 진행되고 있다. 이러한 기술 개발과 더불어 항공기의 경우 공중 운송수단 특성상 사고로 인하여 유발되는 피해가 타 운송수단보다 매우 큰 만큼 사고예방 및 항공안전 향상을 위한 연구도 활발하게 진행되고 있다.

GAMA(General Aviation Manufacturer Association)에서는 미국연방교통안전위원회(NTSB, National Transportation Safety Board)와 미국연방항공청(FAA, Federal Aviation Administration)의 항공사고 데이터를 바탕으로 항공기 사고에 대한 통계자료를 발표하였다[1]. 발표된 통계자료에 따르면 1980년대 이후 30년간 항공기 사고율은 약 40% 정도 감소하였으며, 이는 항공기 설계기술의 비약적 발전과 함께 항공안전에 대한 관심증대로 인한 결과임을 유추해 볼 수 있다. 특히 승객 운송 항공의 경우 사고감소율은 70% 이상으로, 민수 항공여객 분야의 항공안전은 더욱 중요시된 결과임을 확인 할 수 있다. 그러나 전체 항공사고 중 사망 사고 발생률은 1980년대와 2015년 비교 시 여전히 20% 내외이며, 이는 자동차 사망 사고 발생률[2]의 약 20배 이상 높은 수준임을 알 수 있다. 따라서 항공기 사고에 따른 사망을 예방하기 위한 연구의 필요성이 증대되어야 함을 알 수 있다.

국내 많은 기관들에서 항공기 사고에 관한 데이터를 분석하여 그 원인을 발견하고, 문제점 해소를 통하여 항공안전 기술 및 정책 발전에 기여하는 연구들을 진행하고 있다. 그러나 주로 항공기 사고 데이터에 통계분석을 한 연구가 대부분이며, 항공기 사고 데이터의 접근 제한성 및 국내에서 수집 가능한 항공기 사고 데이터의 한계로 인하여 관련 연구가 활발하게 이루어지고 있지 않다. 이에 본 연구에서는 데이터의 다양성 및 접근성이 보장되는 NTSB와 FAA에서 공개하고 있는 약 100,000 여건의 항공기 사고 데이터를 활용하고자 한다. 또한 데이터 마이닝 기법을 적용하여 항공기 사고 및 준사고 때 발생하는 사망 사고에 관한 예측모형을 개발하고, 각 사고의 주요 요인들과 사고 패턴을 분석하는 연구를 진행하고자 하며, 이를 통해 향후 항공기 사고 예방을 위한 실용적인 방안들을 제시함으로써 기존 연구와의 차별성을 가지고자 한다. 향후 본 연구를 통해 국내 항공기 사고와 관련된 여러 연구들에서 데이터 마이닝 기술을 활용한 분석이 활성화되기를 기대해 본다.

본 연구의 구성은 다음과 같다. 2장에서는 항공기 사고와 관련된 주요 용어들을 정의하고 국내외 항공기 안전과 사고 분석에 관한 선행연구들을 살펴보고자 한다. 3장에서는 본 연구에 제시한 연구 프레임워크와 주요 과정들을 소개하고, 4장에서는 연구 프레임워크에 따라 NTSB 데이터와 FAA 데이터를 분석하고자 한다. 5장에서는 각 데이터로부터 항공기 사고 및 준사고 시 사망에 영향을 주는 요인들과 패턴들을 도출하고 그 결과를 비교하고자 한다. 끝으로 6장에서 연구의 시사점과 향후 연구 및 연구 한계점에 대해 기술하고자 한다.

2. 문헌연구

2.1 항공기 사고(Accident) 및 준사고(Incident)

항공기 사고 구분에 대한 국가별 분류에는 일부 차이가 있으나 국제민간항공기구(ICAO, International Civil Aviation Organization)[3]에서 제시한 항공기 사고와 준사고 기준을 보편적으로 사용하고 있다. 국내 항공법 또한 국제민간항공기구에서 제시한 기준(IIS, International Investigation Standards)에 의거하여 항공기 사고 및 항공기 준사고를 다음과 같이 구분하고 있다.

먼저 항공기 사고는 항공기가 이륙하기 이전부터 승객이나 승무원이 항공기에 탑승한 후 내릴 때까지 항공기 운항으로 인한 사람의 사망 또는 부상, 항공기의 손상 등의 항공기와 관련된 모든 사고를 의미한다. 다음으로 항공기 준사고의 경우, 항공기 사고 외에, 항공기 운항과 관련된 안전에 영향을 미칠 수 있는 모든 발생사고로 정의된다.

본 연구에서도 국제민간항공기구에서 제시한 항공기 사고와 항공기 준사고 정의에 따라 NTSB 데이터와 FAA 데이터를 분석하였다.

2.2 연구 동향

국내의 항공기 사고 관련 연구를 요약하면 다음과 같다. 홍승범 외[4]는 유럽에서 개발되고 2010년부터 한국에 도입된 항공사고 데이터 입력 시스템인 ECCAIRS(European Coordination Centre For Aviation Incidents Reporting System)에 입력된 392건의 국내 항공사고 데이터를 활용하여 항공기 유형별, 월별, 비행단계별 및 분류체계별 사고 발생률을 분석하였다. 그 결과 하절기에 항공기 사고가 집중되었으며, 비행기가 순항비행(cruise)을 하거나

착륙(landing)할 때 사고의 발생 빈도가 높았다.

김원규 외[5]는 1991년부터 2010년까지 20년간 군에서 발생한 항공기 사고 사례를 ECCAIRS를 이용하여 분석하였고, 군사 영역에서 ECCAIRS 프로그램의 활용 가능성을 확인하고자 하였다. 그 결과 ECCAIRS의 사고분류체계를 군 목적에 맞게 일부 조정할 경우 ECCAIRS가 군 항공기 사고분석에 활용될 수 있음을 언급하였다.

노태우와 박재찬[6]은 항공사의 안전과 경쟁력 간의 관계에 관한 탐색적 연구를 실시하기 위해 국내 항공기 사례를 대상으로 사고 요인, 사고 예방 및 후속 조치에 대한 분석을 진행하였다. 그 후, 분석 결과를 토대로 국적 항공사의 제도적 측면과 조직 문화적 측면에서 필요한 안전 개선 전략들을 제시하였다.

채정기[7]는 항공기 장애와 관련된 데이터를 대상으로 텍스트 마이닝 분석을 실시하였다. 이를 통해 항공기 장애와 관련된 토픽들과 주요 용어들을 도출하였고 용어들 간의 관계를 통해 항공 사고의 장애요인에 대한 잠재 원인들을 파악하였다.

국외의 항공기 사고 관련 연구를 요약하면 다음과 같다. Lukacova 외[8] 및 Gurbuz 외[9]는 FAA의 준사고 데이터에 분류분석 기법 중 하나인 의사결정나무 알고리즘을 적용하여 사망 예측모형을 개발하고 사망 패턴들을 도출하는 연구를 진행하였다.

Nazeri 외[10]는 데이터 마이닝 기법을 적용하여 각 기관의 항공기 사고와 준사고를 유발하는 주요 요인과 그 차이점에 대해서 분석하였다. 분석결과 항공기 요인(engine, flight control, landing gear)들이 주로 준사고를 유발하였으며, 그 외, 사고와 관련된 항공 운항, 조종사, 회사관련 요인들도 발견하였다.

Christopher와 alias Balamurugan[11]은 여러 분류분석 알고리즘을 적용하여 항공기 사고 예측모형들을 구축한 후, 의사결정나무에서 가장 높은 정확도를 기록한 것을 확인하였다.

그 밖에 통계 분석을 활용하여 헬리콥터의 사고 요인[12]과 일본 항공기의 사고 요인[13]을 분석한 연구들도 확인할 수 있었다.

3. 연구방법

3.1 연구 프레임워크

본 연구에서는 Flg. 1과 같이 연구 프레임워크를 구성하여 연구를 진행하였다. 연구 프레임워크는 데이터 수집, 전처리, 데이터 균형화, 변수 선택, 예측모형 구축, 데이터 분석 및 결과 도출 단계로 구성된다.

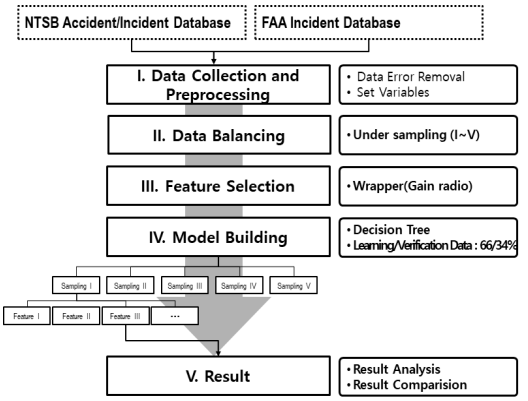

Fig. 1. Research Framework

3.2 데이터 수집

현재 국가별로 다양한 항공 안전정보 관리 시스템을 운영하여 항공기 사고에 관한 데이터를 수집 및 분석하고 있다. 국내의 경우 국토교통부에서 운영하는 통합항공안전정보시스템이 있으며, 국외의 경우 국제민간항공기구의 ADREP(Accident, Incident Data Reporting), FAA의 ASIAS(Aviation Safety Information Analysis & Sharing), NTSB의 항공기 사고 데이터베이스, 미국 항공우주국의 ASRS(Aviation Safety Reporting System), 그리고 유럽의 ECCAIRS(European Coordination Centre for Aviation Incidents Reporting System) 등이 여기에 해당된다.

본 연구에서는 앞서 언급한 여러 항공기 사고 데이터 중 상대적으로 많은 사고 사례를 보유하고 있고, 일반인의 데이터 접근이 용이하며, 정형화된 형태로 분석 데이터를 제공하는 NTSB의 항공기 사고 및 준사고 데이터[14]와 FAA의 항공기 준사고 데이터[15]를 활용하였다.

이를 위해, NTSB에서는 1982년부터 2014년까지 총 31개 변수로 구성된 79,025건의 항공기 사고 및 준사고 데이터를 수집하였고, FAA에서는 1978년부터 2016년까지 총 27개의 변수로 구성된 100,000건의 항공기 준사고 데이터를 수집하였다. 수집된 데이터에 대한 내용은 Table 1과 같다.

Table 1. Description of data sets

Database	NTSB	FAA
Source	Aviation Accident Database	Accident and Incident Data System
	http://www.ntsb.gov/	http://www.asias.faa.gov/
Type of accident	Accident and Incident	Incident
Number of variables	31	27
Year	1982~2014	1978~2016
Amount	79,025	100,000

3.3 전처리(Preprocessing)

지난 30년 동안 항공기 사고의 발생률은 감소하고 있지만 사망률은 크게 감소하지 않기 때문에 본 연구에서는 사망 사고에 대한 요인 분석에 초점을 두고 연구를 진행하고자 한다.

일반적으로 전처리 단계에서는 정보를 정제하는 과정이 포함되는데[16], 본 연구에서는 전처리 과정에서 목표 변수를 변환하고 독립 변수를 제거하는 작업을 진행하였다. 먼저 수집된 NTSB 데이터와 FAA 데이터 중에서 항공기 사망과 관련된 변수를 목표 변수(target variable)로 선정하였다. NTSB에서는 부상 심각도(Injury Severity) 변수를 FAA에서는 총사망률(Total Fatalities) 변수를 목표 변수로 선택하였다. 이때 각각의 목표 변수의 값은 사망 발생 인원을 나타내고 있었다. 일반적으로 항공기 탑승 인원이 많을수록 사망 발생 인원이 많아진다. 본 연구에서는 이와 같은 특성을 분석에서 제거하기 위해 목표 변수의 값을 사망 발생여부인 이진변수로 변환하여 사용하였다.

다음으로 독립 변수 후보군들 중에서 모든 변수의 값이 달라 분석에 불필요하거나 다수의 널(null) 값을 보유하고 있어 분석에 활용하기 어려운 독립 변수들을 제외하였다.

3.4 데이터 균형화(Data Balancing)

학습 데이터(train data)를 통해 예측모형을 구축할 때 목표 변수의 클래스 분포를 우선 파악해야 한다. 이유는 목표 변수 내에 특정 클래스에 속하는 값이 다른 클래스에 속하는 값보다 많을 경우, 일반적으로 학습은 관측 대상이 많은 클래스 위주로 이루어지기 때문에 특정 클래스만을 잘 예측하는 편향된 예측모형이 구축될 수 있기 때문이다. 따라서 예측모형을 구축하기 전에 데이터 균형화를 통해 목표 변수에 존재하는 클래스들의 비율을 맞추는 작업이 필요하다.

본 연구에서 선택한 목표 변수들의 경우 사망 발생보다는 사망 미발생으로 분포가 편향되어 있어서 사망 미발생 클래스를 기준으로 사망 발생 클래스를 언더샘플링(undersampling)하여 클래스의 비율을 맞추는 작업을 진행하였다.

3.5 변수 선택(Feature Selection)

전처리 단계에서 선택된 초기 독립 변수 후보들 중, 목표 변수 예측과 평가에 기여하지 못하는 변수들이 존재할 경우 분석 결과의 정확도를 높이기 위해 해당 변수들을 제거하고 예측모형을 구축하는 것이 바람직하다. 이와 같이 독립 변수 후보들 중 목표 변수 예측에 기여하는 변수들을 선정하는 작업을 변수 선택이라 한다. 본 연구에서는 이득비(gain ratio)를 이용하여 변수들의 중요도를 계산하였으며, 래퍼(wrapper)방식 중 하나인 역방향 제거 기법을 사용하여 예측모형에 포함된 독립 변수를 선택하였다.

3.6 예측모형 구축

본 연구에서는 항공기 사고 및 준사고 때 발생되는 사망여부에 관한 예측모형을 구축하기 위해 데이터 마이닝 툴인 웨카(Weka) 버전 3.8을 이용하였다. 또한 예측모형을 구축할 때 사용될 수 있는 여러 알고리즘들 중에서 의사결정나무(decision tree)를 활용하였다. 그 이유로는 최근 인공지능 분야에서 많이 사용되는 인공신경망의 경우 결과를 해석부분이 블랙박스 형태로 나타나기 때문에 규칙 간의 관계를 파악하기가 어렵지만, 의사결정나무의 경우 분류 규칙들이 이해하기 쉬운 형태로 만들어지고 규칙을 활용한 결과 해석 또한 용이하기 때문이다[17-19].

4. 데이터 분석

4.1 NTSB 데이터 분석

NTSB 데이터로부터 언더샘플링을 실시하여 5개의 표본을 추출하여 클래스의 비율을 맞추는 작업을 진행하였다. 그 결과 표본별 20,524건(사망 발생: 10,262건, 사망 미발생: 10,262건)으로 이루어진 데이터 셋을 최종 분석 데이터 셋으로 결정하였다. 본 연구에서는 추출된 표본들을 66%의 학습데이터와 34%의 검증데이터 비율로 분할하여 실험을 진행하였다.

Table 2는 전처리 과정 이후 NTSB 데이터에서 선정된 16개의 초기 독립 변수와 목표 변수를 보여준다. 변수 선택 과정을 통해 항공기의 파괴정도(Aircraft Damage), 운항 날씨(Weather Condition), 비행단계(Broad Phase of Flight), 엔진수량(Number of Engines), 항공사고 조사 분류 (Investigation Type) 등의 순으로 목표 변수 예측에 많이 기여를 하는 변수들로 분석되었다.

Table 2. Initial variables from NTSB

Type of Variable		Name
Independent Variable	1	Investigation Type
	2	Day
	3	Month
	4	Year
	5	Location
	6	Country
	7	Aircraft Damage
	8	Aircraft Category
	9	Make
	10	Model
	11	Amateur Built
	12	Number of Engines
	13	Engine Type
	14	Purpose of Flight
	15	Weather Condition
	16	Broad Phase of Flight
Target Variable		Injury Severity

각 표본별로 역방향 제거기법을 적용하여 매회 중요도가 낮은 변수를 제거하면서 남은 변수들로 예측모형을 개발하였고, 각 예측모형들의 정확도를 요약하면 Table 3과 같다.

Table 3. Accuracy of Prediction Models in NTSB

No.	Before Sampling		Sample 1		Sample 2	
	RV*	Result (%)	RV	Result (%)	RV	Result (%)
0	All	89.1121	All	84.0499	All	84.0355
1	F2	89.1121	F2	84.2792	F2	84.0355
2	F4	89.1121	F4	84.2792	F4	84.6088
3	F3	89.1121	F3	84.2792	F3	84.0355
4	F5	89.1121	F8	84.2792	F5	84.0355
5	F8	89.1121	F5	84.2792	F8	84.0355
6	F11	89.1121	F11	84.2792	F11	84.0355
7	F13	89.1121	F13	84.2792	F13	84.0355
8	F14	88.4343	F14	84.1215	F14	84.0355
9	F9	88.4343	F9	84.0355	F9	83.9496
10	F10	89.4902	F12	83.4766	F12	83.8206
11	F12	89.3920	F10	83.8779	F10	83.8206
12	F16	88.4343	F6	83.8779	F6	83.8206
13	F1	88.4343	F16	83.4623	F16	83.4623
14	F6	88.4343	F1	83.5053	F1	83.4623
15	F15	88.4343	F15	83.5196	F15	83.5483
16	F7	–	F7	–	F7	–

No.	Sample 3		Sample 4		Sample 5	
	RV	Result (%)	RV	Result (%)	RV	Result (%)
0	All	85.0817	All	83.6199	All	83.5483
1	F2	85.0817	F2	83.7059	F2	83.5483
2	F4	85.0817	F4	83.7059	F4	83.5483
3	F3	84.4941	F3	83.7059	F3	83.5483
4	F5	84.4941	F5	83.7059	F5	83.5483
5	F11	84.4941	F8	83.7059	F8	83.5483
6	F8	84.4941	F11	83.7059	F13	83.5483
7	F13	84.4941	F13	83.7059	F11	83.5483
8	F14	84.4081	F14	83.7059	F14	83.5483
9	F12	83.8779	F9	83.6486	F9	83.7919
10	F9	83.8779	F12	83.1614	F12	83.2330
11	F10	84.3078	F10	83.6056	F10	83.7203
12	F6	84.3078	F6	83.6056	F6	83.7203
13	F16	83.9782	F16	83.4910	F16	83.7489
14	F1	83.8922	F1	83.4336	F1	83.7489
15	F15	83.7919	F15	83.1040	F15	83.2330
16	F7	–	F7	–	F7	–

* RV: Removed variable

Table 3을 보면, 언더샘플링을 하지 않고 전체 NTSB 데이터를 통해 예측모형을 구축한 경우가 표본에서 예측모형을 구축한 것보다 일반적으로 높은 정확도를 기록한 것을 확인할 수 있다. 그러나 이는 앞서 설명한 바와 같이 편향된 목표 변수의 값으로 인한 결과임을 알 수 있다. NTSB 데이터에서 가장 높은 정확도인 85.0817%를 보여준 예측모형은 표본 3에서 14개의 변수로 예측모형을 구축하였을 때이다. 본 연구에서는 해당 예측모형을 통해 사망관련 주요 패턴들을 도출하고자 한다.

4.2 FAA 데이터 분석

NTSB 데이터 분석과 동일하게 FAA 데이터에서도 언더샘플링을 실시하여 5개의 표본을 추출하였고 표본별 1,566건(사망 발생: 783건, 사망 미발생: 783건)으로 균형화한 데이터 셋을 최종 분석 데이터 셋으로 사용하였다. FAA 데이터의 경우 준사고 데이터만을 포함하고 있기 때문에 사고와 준사고를 모두 포함하고 있는 NTSB 데이터 보다는 사망 발생 건수가 적음을 알 수 있다.

Table 4. Initial variables from FAA

Type of Variable		Name
Independent Variable	1	Local Event Date
	2	Event State
	3	Aircraft Damage
	4	Flight Phase
	5	Aircraft Make
	6	Aircraft Model

Type of Variable		Name
	7	Flight Conduct Code
	8	Total Injuries
	9	PIC Certificate Type
	10	PIC Flight Time Total Hrs
	11	PIC Flight Time Total Make Model
Target Variable		Total Fatalities

Table 4는 FAA 데이터에서 선정된 11개의 초기 독립 변수와 목표 변수를 보여준다. 변수 선택과정을 통해 독립 변수들의 중요도를 계산한 결과 NTSB 데이터와 유사하게 항공기의 파괴정도(Aircraft Damage)와 비행단계(Flight Phase)가 목표 변수에 영향을 크게 미치는 독립 변수로 분석되었다. NTSB 데이터에 없는 중요 변수로는 항공기의 운항목적(Flight Conduct Code)과 조종사의 능력을 나타내는 조종사 인가여부(PIC Certificate Type), 조종사 총 비행시간(PIC Flight Time Total Hrs), 조종사 해당 기종 비행시간(PIC Flight Time Total Make Model) 변수가 분석되었다.

각 표본별로 역방향 제거기법을 사용하여 매회 중요도가 낮은 변수를 제거하면서 남은 변수들로 예측모형을 개발하였고, Table 5와 같이 각 예측모형들의 정확도를 요약하였다.

Table. 5 Accuracy of Prediction Models in FAA

No.	Before Sampling		Sample 1		Sample 2	
	RV	Result (%)	RV	Result (%)	RV	Result (%)
0	All	99.7308	All	96.9925	All	96.9925
1	F2	99.7308	F2	96.9925	F2	96.9925
2	F12	99.7308	F12	96.9925	F12	96.9925
3	F1	99.7308	F1	96.9925	F1	96.9925
4	F6	99.7308	F5	96.9925	F5	96.9925
5	F5	99.7358	F6	96.9925	F6	96.9925
6	F11	99.7607	F11	97.1805	F11	97.1805
7	F4	99.7358	F10	96.6165	F9	96.6165
8	F9	99.7408	F9	96.6165	F10	96.6165
9	F10	99.6061	F4	96.4286	F4	96.2406
10	F7	98.6489	F7	96.4286	F7	96.0526
11	F3	-	F3	-	F3	-
No.	Sample 3		Sample 4		Sample 5	
	RV	Result (%)	RV	Result (%)	RV	Result (%)
0	All	94.5489	All	96.0526	All	94.9248
1	F2	94.5489	F2	96.0526	F2	94.9248
2	F12	94.5489	F12	96.0526	F12	94.9248
3	F1	94.5489	F1	96.0526	F1	94.9248
4	F6	93.9850	F5	96.0526	F5	94.9248
5	F5	95.4887	F6	95.6767	F6	94.9248
6	F11	95.4887	F11	95.8647	F11	94.9248
7	F9	94.7368	F9	94.5489	F9	94.9248
8	F10	94.7368	F10	94.5489	F10	94.9248
9	F4	94.7368	F4	94.9248	F4	94.3609
10	F7	94.7368	F7	94.9248	F7	94.3609
11	F3	-	F3	-	F3	-

Table 5를 통해 가장 높은 정확도인 97.1805%를 보인 예측모형은 표본 1에서 5개의 변수로 예측모형을 구축하였을 때임을 알 수 있으며, 본 연구에서는 이 해당 예측모형을 바탕으로 FAA 데이터에 대한 사망관련 주요 패턴들을 도출하고자 한다. 전반적으로 Table 5에 나타난 FAA 데이터의 예측 정확도는 Table 3에 나타난 NTSB 데이터의 예측 정확도보다 약 10% 이상 높은 수치를 보여주고 있다. 이는 상대적으로 특정 상황에서만 국한되어 발생하는 항공기 준사고의 특성에 의한 결과임을 확인할 수 있다.

5. 결과 도출

5.1 NTSB 데이터의 결과 도출

본 절에서는 NTSB 데이터에서 언더샘플링한 표본들 중 가장 좋은 성능을 보여준 표본 3의 의사결정나무를 활용하여 항공기 사고 및 준사고 때 발생되는 사망여부에 대한 주요 패턴들을 도출하고 사망 사고 예방을 위한 방안들을 제시하고자 한다.

Fig. 2는 의사결정나무에서 도출된 15개의 패턴들을 보여주며, 7개의 사망패턴(Fatal)과 8개의 비사망 패턴(Non Fatal)을 확인할 수 있다. 발견된 패턴의 주요 내용을 요약하면 다음과 같다.

항공기 사고 및 준사고로 인하여 사망이 발생하는 가장 큰 패턴은 항공기가 사고 후 완파되었을 경우(Aircraft Damage = Destroyed)로 약 91% 수준의 사망이 발생함을 알 수 있다. 이러한 이유로 항공기 내추락 시 안전성 확보를 위한 연구와 함께 내추락 관련 감항요구사항은 지속적으로 항공기 개발간 주요 고려사항으로 요구되고 있으나, 상대적으로 비행간 항공기 추락사고의 경우 기체가 완전히 파괴될 가능성이 높기 때문에 단순히 내추락 성능만을 향상시키는 설계 노력만으로는 추락사고 발생에 따른 생존률 보장에 한계가 있다. 또한 내추락 성능의 향상은 항공기의 중량/비용 증가를 반드시 수반함을 인지할 필요가 있다. 따라서 개발초기 단계에서부터 철저하게 항공기의 주요 임무와 운용 환경을 고려하여 적절한 내추락 설계수준을 정하는 것도 항공기 안전성과 함께 운용효과 측면에서 중요한 고려요소중 하나이다.

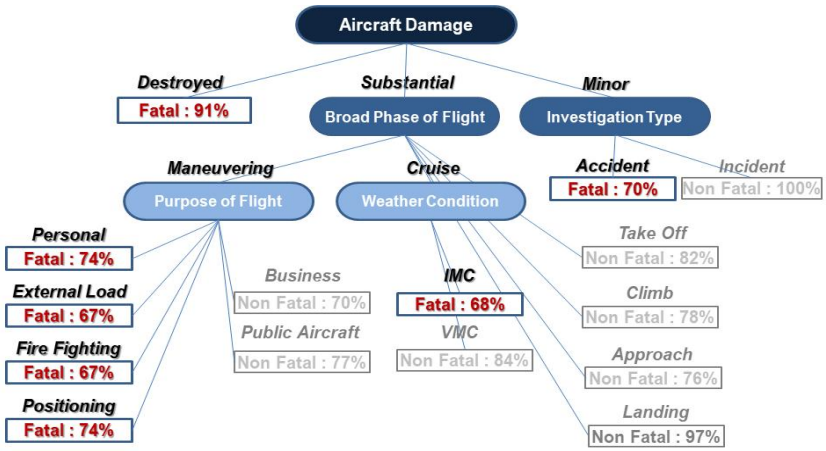

Fig. 2. Decision Tree of NTSB

다음으로는 항공기가 일부 파괴되었을 경우(Aircraft Damage = Substantial)에는 주로 고기동 비행단계(Broad Phase of Flight = Maneuvering) 또는 외부화물공수(Purpose of Flight = External Load), 화재진압(Purpose of Flight = Fire Fighting) 등과 같은 고위험 임무를 수행할 때 약 67% 이상의 사망이 발생함을 알 수 있다. 따라서 고기동, 고위험 임무수행에서 발생되는 항공사고를 예방을 위하여 항공기 기동성 향상, 자동비행조종기능부여 및 임무별 적합한설계(mission dependent design)와 같은 노력이 필요하며, 적합한 임무장비 부여 및 철저한 비행운영관리(숙련된 조종사, 예방정비 강화) 방안 또한 요구된다.

마지막으로 항공기 순항비행 단계에서(Broad Phase of Flight = Cruise) 날씨가 좋지 않을 경우(Weather Condition = IMC)에는 약 68% 수준의 사망이 발생하였음을 알 수 있다. 여기에서 IMC(Instrument Meterological Condition)는 악천 후로 인하여 시계비행이 불가능할 경우 계기를 통한 비행이 필요한 조건을 나타낸다. 따라서 이와 같은 악천후 기상조건에서의 비행안전성 향상을 위하여, 다양한 고성능 항법장비 및 항법보조장비(GPS/INS, VOR/ILS, 3D Digital Map) 등을 항공기에 장착하는 방안이 마련되어야 한다.

5.2 FAA 데이터의 결과 도출

Fig. 3은 FAA 데이터에서 가장 높은 성능을 나타낸 표본 1의 의사결정나무를 보여주고 있다. 의사결정나무를 통해 항공기 준사고 때 발생되는 6개의 사망 패턴과 9개

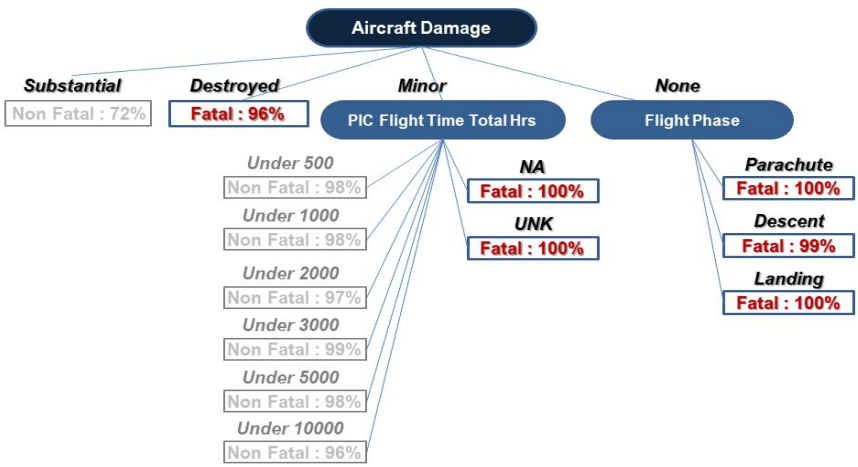

Fig. 3. Decision Tree of FAA

의 비사망 패턴을 도출하였으며 주요 내용은 다음과 같다.

앞서 살펴본 NTSB 데이터에서 발생한 패턴과 동일하게 항공기 준사고로 인하여 사망이 발생하는 패턴에는 항공기가 사고 후 완파되었을 경우(Aircraft Damage = Destroyed)로 약 96% 수준의 사망이 발생함을 알 수 있었다. 그러나 FAA 데이터에서는 항공기의 준사고만을 다루다 보니 전체 사고 783건 중 항공기가 완파된 경우가 20건 밖에 없어 항공기 파괴여부는 다른 요인들과 비교하여 준사고 시 사망발생에 큰 영향을 미치는 요인이 아님을 알 수 있다.

FAA 데이터의 경우 항공기가 파괴가 되지 않았을 경우(Aircraft Damage = None)에도 다수의 사망이 발생한 것을 확인할 수 있었다. 이는 주로 고공낙하점프(Flight Phase = Parachute Jumping)로 인한 사망 발생의 경우가 대다수(500여건)를 차지하였으며, 그밖에 항공기 하강(Flight Phase = Decent) 및 착륙(Flight Phase = Landing Approach) 단계에서도 일부 사망이 발생되는 것을 알 수 있었다. FAA 데이터에서 발생된 준사고의 경우 항공기 운항 시 발생하는 기체관련 사고 외에 다양한 사망 패턴들(고공낙하로 인한 사망, 비행간 탑승객의 심장마비, 비행기 하차 시 승객 계단 추락 등)을 포함하는 만큼 특정 패턴에만 적용할 수 있는 예방방안보다는 사고별 발생 상황과 상세 원인분석을 통한 세부 예방방안의 마련이 필요하다.

5.3 결과 비교

본 연구에서는 NTSB 데이터와 FAA 데이터를 대상으로 항공기 사고와 준사고로 인한 사망 요인들과 패턴들을 분석하는 연구를 실시하였다. 두 기관 모두 미국 정부의 항공관련 기관임에 따라 항공기 사고 분류와 사고 관련 제반 기준들은 동일하다고 볼 수 있다. 그러나 제공되는 데이터들에는 다음과 같은 차이점이 존재하였다.

첫째, NTSB 데이터는 항공기 사고 및 준사고 데이터를 모두 제공하고 있으나, FAA 데이터의 경우 항공기 준사고 데이터만을 제시하고 있다.

둘째, 각 기관에서 제공하는 항공 사고 데이터의 양은 NTSB에서는 약 79,000건, FAA에서는 약 100,000건 정도로 비슷한 수준이었으나, 언더샘플링 후 분석에 활용된 데이터의 경우 NTSB에서는 20,524건, FAA에서는 1,566건으로 데이터의 양에 있어서 크게 차이를 보였다.

셋째, 각 기관에서 제공된 데이터 모두 항공기 파괴여부, 비행정보, 사고항공기, 제조사, 엔진정보, 사망여부 등과 같은 공통속성을 보유하고 있음과 동시에, 각 기관에서만 제공하는 고유의 속성들도 존재하였다. NTSB 데이터의 경우 항공기 종류와 날씨 정보가 포함되어 있었으며, FAA 데이터에서는 조종사의 능력과 관련된 변수들을 보유하고 있었다.

다음으로, NTSB 데이터와 FAA 데이터를 통해 구축된 의사결정나무 기반 예측모형들의 특성과 각 모형에서 도출한 항공기 사고 및 준사고로 인한 주요 사망 요인들과 패턴들을 비교하였다.

첫째, 최고 정확도를 기록한 예측모형에 있어서 NTSB 데이터에서는 85.08%, FAA 데이터에서는 97.18%를 기록하여 FAA 데이터를 활용한 예측모형이 상대적으로 높은 정확도를 보여주었다. FAA 데이터의 경우 항공기 준사고만을 다루고 있기 때문에 그 특성이 반영된 결과로 해석된다. 항공기 준사고의 경우 사고분류 특성상 사망 발생건수가 783건으로 상대적으로 미비하고, 사망을 유발하는 조건도 상대적으로 다양하지 않은 결과로 분석된다.

둘째, 항공기 사망에 영향을 주는 요인들 중 공통 요인으로는 항공기 파괴여부와 비행특성이 식별되었으며, 차별 요인으로는 NTSB 데이터의 경우 날씨 관련 요인, FAA의 경우 조종사의 능력과 관련 요인들이 확인되었다.

셋째, 최종 예측모형을 통한 도출된 패턴을 비교해 보면, NTSB의 경우 항공기의 완파, 고기동 상태에서 위험임무 수행, 순항비행 상태에서의 악천후 날씨가 사망을 유발하는 주요 패턴으로 분석되었다. FAA의 경우도 항공기 완파가 사망을 유발하는 주요 패턴으로 식별되었으나 전체 사고와 비교했을 때 그 건수가 미비하였다. 대신 항공기가 일부 파괴된 경우 조종사의 숙련도가 저조하거나 미인가 조종사의 경우 사망이 자주 발생하였으며, 준사고 데이터에서 가장 많은 사망 사례를 기록한 고공낙하점프와 지상운용단계에서 발생되는 여러 사망 사고와 관련된 패턴들을 발견할 수 있었다.

6. 결론

본 연구에서는 NTSB 데이터와 FAA 데이터모두를 대상으로 데이터 마이닝 기법을 적용하여 항공기의 사고 및 준사고 때 발생되는 사망여부를 예측하는 모형들을 구축한 점, 구축된 예측모형을 통하여 사망에 영향을 주는 주요 요인들과 패턴들을 도출한 점, 그리고 패턴 분석을 통해 항공기 사망 사고 예방을 위한 방안들을 제시한 점에서 기존 연구와의 차별성과 학문적 시사점을 찾을 수 있다.

본 연구에서 도출한 주요 요인들과 패턴들은 향후 실무에서 항공기 사고 예방과 항공안전 향상을 위한 활동들을 계획할 때 활용될 수 있을 것이다.

데이터 마이닝 분석의 경우 수집된 데이터의 특성이 분석 결과에 많은 영향을 주게 된다. 본 연구에서는 항공기 사고와 관련하여 각 기관에서 제공되는 정형 데이터만을 활용하여 분석을 진행하였다. 현재 각 국의 감항당국은 항공기 사고에 관한 많은 양의 사고분석 보고서를 텍스트 형태로 보유하고 있으며 그 일부를 공유하고 있다. 따라서 예측모형을 구축할 때 해당 보고서에 텍스트 형태로 명시되어 있는 비정형 데이터들을 활용하는 연구를 추가로 진행한다면 보다 다양한 연구 결과들이 도출될 것을 기대해 볼 수 있다. 또한 FAA 데이터에서 나타난 항공기 준사고의 경우 적은 수의 사망 사고가 특정 요인들에 편중되어 발생하였기 때문에 준사고로부터 다양한 사망관련 패턴을 도출하지 못한 점은 본 연구의 한계점으로 인식된다.

REFERENCES

[1] GAMA. (2016). 2016 General Aviation Statistical Databook & 2017 Industry Outlook, America, General Aviation Manufacturers Association.

[2] GOV. UK. (2015). *Number of fatalities resulting from road accidents in Great Britain table.* https://www.gov.uk/government/publications/annual-road-fatalities

[3] ICAO. *International Investigation Standards.* http://www.icao.org/icao/en/cat.htm

[4] S. B. Hong, W. Y. Kim & Y. C. Choi. (2012). The Trend Analysis about Aviation Accident and Incident in Korea Using the ECCAIRS Data. *The journal of Korea Navigation Institute, 16(4),* 687-696.

[5] W. Kim, S. Hong, M. Jie, G. Hong, D. Ahn & C. Choi. (2013). Analysis of Aviation Accident and Incident in Military Using the ECCAIRS 5. *Journal of the Korean Society for Aviation and Aeronautics, 21(1),* 80-86.

[6] T. Roh & J. Park. (2014). An Explanatory Study on Air Accident Analyses and Strategy for Improvement in Air Safety. *Journal of the Aviation Management Society of Korea, 12(4),* 95-124.

[7] J. Chae. (2014). *A study of the application of Big data analytics in Aviation Safety*, master dissertation, Ewha Womans University, Seoul.

[8] A. Lukacova, F. Babic & J. Parali. (2014). Building the Prediction Model from Aviation Incident Data. *IEEE 12th International Symposium on Applied Machine Intelligence and Information,* 365-369.

[9] F. Gurbuz, L. Ozbakir & H. Yapici. (2009). Classification rule discovery for the aviation incidents resulted in fatality. *Knowledge-based system, 22(8),* 622-632.

[10] Z. Nareri, G. Donohue & L. Sherry. (2008). Analyzing Relationships Between Aircraft Accidents and Incidents - A data Mining Approach. *In proceeding of Third International Conference on Research in Air Transportation, Virginia, US*, 185-190.

[11] A. A. Christopher & S. A. alias Balamurugan. (2014). Prediction of warning level in aircraft accidents using data mining techniques. *The Aeronautical Journal, 118(1206),* 935-951.

[12] A. D. Voogt & R. A. V. Doorn. (2007). Helicopter accident: Data-mining the NTSB database. *European Rotorcraft Forum 33.*

[13] K. Iwadare. (2015). Statistical Data Analyses on Aircraft Accidents in Japan: Occurrences, Causes and Countermeasures. *American Journal of Operations Research, 5(3),* 222-245.

[14] NTSB. *Aviation Accident Database.* http:// www.ntsb.gov/

[15] FAA. *Accident and Incident Data System.* http://www.asias.faa.gov/

[16] J. Lee. (2014). A Study on the Data Mining Preprocessing Tool For Efficient Database Marketing. *Journal of Digital Convergence, 12(11),* 257-264.

[17] J. Oh & S. Choi. (2018). An Analysis of the Characteristics of Companies Introducing Smart Factory System Using Data Mining Technique. *Journal of the Korea Convergence Society, 9(5),* 179-189.

[18] J. Lee & H. Lee. (2018). Meltdown Threat Dynamic Detection Mechanism using Decision- Tree based Machine Learning Method. *Journal of Convergence for Information Technology, 8(6),* 209-215.

[19] M. Park. (2018). Determinant of the Elderly Poverty Using Decision Tree Analysis. *Journal of Digital Convergence, 16(7),* 63-69.

김 정 훈(Jeong-Hun Kim) [정회원]

· 2007년 2월 : 한국항공대학교 기계공학 (학사)
· 2010년 2월 : 경상대학교 항공우주공학 (공학석사)
· 2018년 2월 : 경상대학교 경영대학원 (경영학석사)
· 2006년 9월 ~ 현재 : 한국항공우주산업 개발본부 책임연구원
· 관심분야 : 항공사고 분석, 데이터 마이닝
· E-Mail : kjh0513@koreaaero.com

김 태 운(Tae-Un Kim) [학생회원]

· 2013년 3월 ~ 현재 : 경상대학교 경영정보학과 학부생
· 관심분야 : 데이터 마이닝, 빅데이터 분석
· E-Mail : dmldz@naver.com

유 동 희(Dong-Hee Yoo) [정회원]

· 2002년 8월 : 고려대학교 MIS (경영학사)
· 2009년 2월 : 고려대학교 일반대학원 경영학과 MS/IS (경영학박사)
· 2009년 6월 ~ 2013년 5월 : 육군사관학교 전자정보학과 조교수
· 2015년 3월 ~ 현재 : 경상대학교 경영정보학과 부교수 및 경영경제연구소 책임연구원
· 관심분야 : 빅데이터 분석, 인공지능, 지능형 웹
· E-Mail : dhyoo@gnu.ac.kr