머신러닝을 활용한 선박 사고 예측 및 안전 항해 구역 시각화 시스템

안동준*, 김윤지*, 이태검*, 이승수*, 김동재*, 박수현**
동국대학교 정보통신공학과*, ICT 한이음 멘토**
adj0707@gmail.com*, wendyunji1209@gmail.com*, taegeom0601@gmail.com*,
seungsoo0701@gmail.com*, kinbell19@gmail.com*, lovehyun@hotmail.com**

Ship Accident Prediction & Safety territory virtualization System with Artificial intelligence

Dong-jun An*, Yun-ji Kim*, Tae-geom Lee*, Seung-soo Lee*, Dong-jae Kim*, Su-hyun Park**
*Dept. of Information Communication Engineering, **Dong-Guk University, ICT Hanium Mentor

요 약

다수의 사고가 발생되는 소형 선박에 반해 대형 선박을 위주로 제공되고 있는 스마트 해상 물류 시스템을 뒷받침하기 위하여 소형 선박에서 자주 발생할 수 있는 사고의 유형과 그 예상 확률을 제공하는 시스템을 연구하고 제공한다. 로지스틱 분류를 통해 사고의 확률을 예측하며 추천 알고리즘을 활용한 발생 가능성이 높은 사고의 유형을 도출하여 소형 선박용 e-navigation 을 제공한다.

1. 서론

정부에서 해양 수산 분야 한국판 뉴딜 시리즈로 '스마트 항만', '자율 운항 선박' 등을 제시하며 해상 물류에도 도래하고 있다. 2025 년까지 '스마트 해상 물류 체계' 를 구축하고자 하며 자율 운항 선박- 초고속 해상 통신망 등을 연계한 최적화된 해상 물류 체계를 그리고 있다[1]. 실제로 2016 년부터 6 개의 한국형 e-navigation 을 개발하여 전 연안의 선박을 대상으로 다양한 서비스를 제공 중이다. 하지만 이러한 서비스는 대형 선박을 대상으로 운영이 되고 있으며 사고율이 높은 소형 선박의 경우 실정이 달라진 바가 크지 않아 해상 전반의 관리는 어려움을 겪고 있다. 소형 선박 및 어선의 사고율을 감소시킬 경우 더욱더 체계화된 바닷길을 여는 포문이 될 것이다.

따라서 본 논문에서는 선박사고 감소 및 스마트 해상 체계의 기반을 다지기 위한 초석으로 인공지능을 활용한 해양 사고 예측 및 안전 항해 구역 시각화' 시뮬레이터를 설계하고 제작하여 시험한다. 공공데이터로 제공되고 있는 기존의 사고 발생 이력들을 빅데이터로 변환한다. 이후 선박의 사고 확률 및 유형을 예측하고 위험 등급을 산출하여 하나의 MAP 으로 제공한다. 이는 발생할 수 있는 사고를 예측할 수 있다는 측면에서 해양 사고를 대비하고 사고 발생을 감소시킬 수 있으며 추후 제공될 선박 자율주행 기술의 기반으로의 활용 가능성이 높다[2].

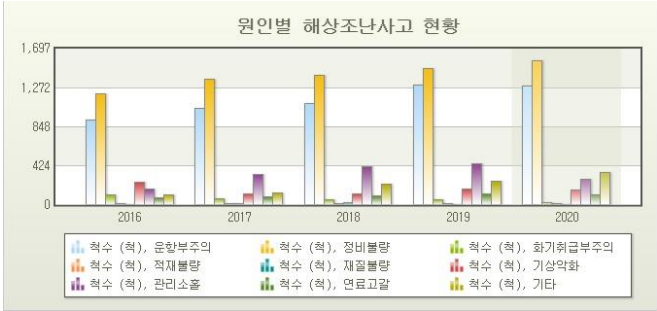
2. 로지스틱 분류를 활용한 사고 확률 예측

머신 러닝은 지도 학습과 비지도 학습으로 분류가 가능하다. 해당 프로젝트에서는 지도 학습을 사용한다. 그 이유는 사고라는 결과 데이터를 바탕으로 그 원인이 기재되어 있는 과거의 데이터가 존재하기 때문이다. 과거의 데이터와 현재 상황의 유사성을 비교하기 위해 지도학습을 선정하였다.

또한 본 연구에서는 사고의 확률이라는 특정한 퍼센트로 그 결과값이 도출이 되어야 하기 때문에 0 과 1 사이의 값을 얻을 수 있는 다항 로지스틱 회귀 방식을 사용하였다.

2.1 데이터 셋의 구성

데이터 셋을 선정하기 앞서 사고의 영향을 주는 요인을 판단하기 위해 다양한 해양 수산부의 보고를 조사하였다. 그 결과 '기상 환경이 사고의 요인이 될 것이다' 라는 가설을 세우게 되었다[4]. 이러한 가설을 바탕으로 CDA 분석을 진행하였고 국가 통계 포털에서 제공하는 해양 사고 및 조난사고의 통계(현황)를 통해서 가장 많은 영향을 주는 사고의 원인은 개인의 부주의이며 그 다음으로는 기상악화의 영향을 받고 있음을 알 수 있었다.



(표 1) 2016~2020 원인 별 해상조난사고 현황

이를 통해 사고 데이터를 기반으로 사고가 자주 발생한 위치와 기상 상황일 경우 높은 사고 확률을 보일 것이라 판단하였다.

활용하는 사고 데이터는 해양수산부 중앙 해양 안전 심판원에서 제공하는 공공 데이터로 2016 년부터 2020 년까지의 해양 사고 데이터를 포함하고 있다. 또한 기상 자료 개방 포털에서 제공하는 기상 데이터를 사용하였으며 날짜, 위도와 경도, 해구의 번호, 유의파고와 파향, 최대 파 주기, 풍속과 풍향으로 구성되어 있다. 이 중, 사고 데이터와 동일한 날짜, 위도 경도를 가지는 데이터들을 추출하여 기존에 가지고 있던 사고 데이터에 유의 파고, 풍속 등의 기상 데이터를 추가하였다

	date	위도	경도	해구번호	유의파고	파향	최대파주기	풍속	풍향
0	2016-05-12 15:00:00	36.25	125.75	173	1.0	203.0	4.5	6.2	183.0
1	2016-03-22 00:00:00	30.75	126.75	528	1.8	38.0	6.3	9.0	43.0
2	2016-11-12 18:00:00	34.75	126.25	204	0.1	192.0	4.1	5.4	129.0
3	2017-05-01 06:00:00	36.75	126.25	164	0.4	236.0	5.6	1.8	166.0
4	2016-05-22 18:00:00	36.25	126.25	174	0.0	213.0	6.3	2.8	325.0

(표 2) 기상데이터를 포함한 데이터 셋 개형

2.2 데이터 전처리(Data Processing)

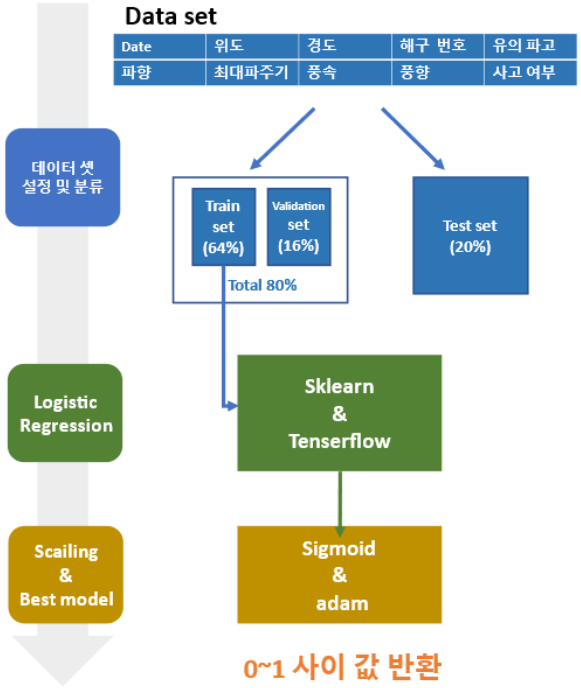
사고의 확률을 도출하기 위해서는 사고가 발생한 데이터와 무사고 데이터로의 분류 과정이 요구된다. 무사고를 표본이 한쪽으로 편향되지 않도록 랜덤으로 열을 추가하여 구성하였다. 이 후 사고의 유무를 1 과 0 으로 판단하는 하나의 행을 추가하여 새로운 데이터(상황)가(이) 주어진 경우 해당 데이터가 가지는 값을 0 과 1 사이의 소수로 반환할 수 있도록 구성하였다.

총 데이터 셋에서 8:2 로 랜덤으로 분류한 후 20% 에 해당하는 사고들을 test set 으로 설정하였다. 이는 만들어진 best model 의 정확도 검사용으로 활용하였다. 또한 Training 과정에서 학습을 담당할 train data 와 중간 평가를 담당할 validation data 로 분류하였다. 또 다시 8:2 의 비율로 train set 과 validation set 을 구성하였으며 이를 통해 best model 을 생성하였다.

2.3 모델의 구성

과거의 사고 데이터를 훈련 데이터로 설정하고 sklearn 과 tenserflow 를 활용하여 머신 러닝을 시킨다. 이를 통해 사고 예측 데이터를 얻을 수 있게 되는데 이는 선형 회귀 방식을 따른다. 이렇게 얻어진 예측 데이터 결괏값에 0~1 사이의 값을 반환하는 sigmoid 연결 함수를 곱하여 스케일링 한다. 이러한 과정을 통해 로지스틱 선형 회귀를 구현하였다.

0~100%의 100 가지로 분류하기 위해서는 다항 로지스틱 회귀를 따라야 한다. 단순 선형 회귀 방식은 하나의 예외 데이터로 인하여 기울기가 변화될 가능성이 높지만 로지스틱 회귀의 경우 0~1 사이의 값을 반환해야 하기 때문에 그 값의 범위가 줄어들어 예외 데이터의 영향을 덜 받을 수 있다. 또한 통상적인 로지스틱 회귀는 적은 손실일 때 분류를 하지만 해당 프로젝트에서는 분류를 내리기 전 실숫값을 사용하였다.



(그림 1) 확률 모델 flowchart

본 연구에서는 Sequential 함수를 통하여 레이어를 선형으로 연결 구성하였으며 은닉층에서 sigmoid 활성화(연결) 함수를 사용하였다. 해당 모델의 경우 사고와 무사고 이진 분류를 거치기 때문에 binary cross entropy 기법을 사용하였으며 데이터의 양이나 은닉층이 많아질수록 시간이 오래 소요되는 점을 고려하여 경사 하강법을 이용하였다. 또한 sigmoid 를 보안하기 위해 momentum 과 adagrad 의 단점을 보완한 RMSProp 을 합친 Adam 을 옵티마이저로 설정하였다. Adam 은 경사하강법을 따르며 best model 을 찾기 위한 방법으로 사용되었다

2.4 정확도 분석

epoch	loss	acc	val_loss	val_acc	epoch	loss	acc	val_loss	val_acc
1	0.5886	0.7111	.	.	11	0.5886	0.7109	0.6018	0.7011
2	0.5886	0.7110	0.6009	0.7018	12	0.5886	0.7108	0.6006	0.7023
3	0.5886	0.7109	0.6006	0.7011	13	0.5887	0.7111	0.602	0.7023
4	0.5886	0.7109	0.6009	0.7011	14	0.5886	0.7103	0.6018	0.7021
5	0.5889	0.7109	0.6018	0.7023	15	0.5886	0.7110	0.6009	0.7023
6	0.5889	0.7107	0.6018	0.7011	16	0.5889	0.7112	0.6018	0.7021
7	0.5888	0.7103	0.6006	0.7023	17	0.5886	0.7113	0.6006	0.7011
8	0.5887	0.7111	0.6018	0.7023	18	0.5887	0.7109	0.6009	0.7021
9	0.5887	0.7110	0.6009	0.7011	19	0.5887	0.7000	0.6009	0.7011
10	0.5888	0.7109	0.6018	0.7023	20	0.5886	0.7103	0.6018	0.7023
최종 best model loss: 0.6018 - acc: 0.7023									

(표 3) 확률 모델 정확도 분석

Epoch(실행 횟수)를 총 20 회로 설정하고 모델을 피팅 해본 결과 loss 값은 0.6018, 정확도는 0.7023 으로 약 70%의 정확도를 얻을 수 있었다.

정확도 개선을 위해 조사한 결과 대한민국 국내 출항 현황이 549615 임에 근거하여 사고 확률은 약 2% 가량임을 알 수 있었다. 따라서 사고:무사고의 비율을 1:50 으로 조정하여 학습한 결과 97%의 정확도를 얻을 수 있었다. 하지만 이는 다수의 경우를 무사고로 판단하여 생기는 과적합 현상으로 판단하여 실질적인 서비스 제공에 어려움이 있다.

3. 추천 알고리즘을 활용한 사고 유형 예측

Recommendation System 은 개인에게 적합한 상품 및 서비스를 제공하는 기술로서 최근 OTT 서비스에서 많이 활용되고 있다. 해당 프로젝트에서 얻고자 하는 결과는 '특정 상황에서 발생 가능한 사고 유형'이다. 선박 출항 전 해당 선박에 발생 가능한 사고 유형을 미리 알고 대비 가능한 경우 해양 사고의 감소를 불러올 것이다. 따라서 사고의 발생 원인에 따른 사고의 유형을 도출해야 하였다. 개인의 선박 종류, 활동 영해, 시간 등에 따라 달라지는 상황에 맞추어 가장 일어날 가능성이 높은 사고의 유형을 도출한다.

이는 사용자의 정보를 기반으로 그 특성을 파악하여 서비스를 추천해주는 것과 유사하게 사용자의 상황을 바탕으로 가장 발생확률이 높은 사고의 유형을 추천해줄 수 있기에 추천 알고리즘을 활용하였다.

3.1 데이터 셋의 구성

기본적인 데이터 셋은 확률 예측과 동일하게 2016~2020 년 제공되는 15208 개의 해양 사고데이터를 활용하며 이를 바탕으로 사고의 유형을 결정하는 요소들이 무엇인지 판단하기 위해 EDA 분석을 진행하였다.

사고 데이터에서 제공하는 사고의 유형은 15 가지로 구분 되어있다. 기관손상, 부유물 감김, 안전사고, 운항 저해, 전복, 접촉, 조타장치손상, 좌초, 추진축계 손상, 충돌, 침몰, 침수, 해양오염, 화재 폭발이며 모든 사고는 하나의 사고로 분류되어 있다. 가장 많이 발생한 사고의 종류는 기관 손상이며 15208 회의 사고 중 4292 회를 차지하였고 충돌 사고가 2553 회로 그 뒤를 이었다.

또한 사고의 변수는 크게 사고의 발생 시점(년도, 월, 일, 시간)과 장소(발생 지역과 정확한 위도, 경도) 그리고 선박과 관련한 정보(선박 명, 톤 수, 용도)이다. 각 정보는 통계적으로 인용 가능한 수치로 분류되어 제공된다. 시간대는 4 시간씩 6 가지, 지역은 영해와 공해, 항구 등을 통해 8 가지, 톤 수 범위는 1 톤 미만부터 100,000 톤 이상까지 총 16 가지, 선박의 용도는 어선, 여객선 등 7 가지로 분류되어 있다.

이러한 데이터들과 얻고자 하는 결과(사고 유형)의 상관 관계를 따져 보기로 하였다. 이를 위해 각 사고의 유형 별 변수들이 가지는 사고의 횟수 중 가장 빈도가 높은 값을 추출하고, 유의미한 값을 가지는 요소들을 다음의 표로 정리하였다.

사고 유형	월	시간대(시)	발생 지역	톤 수 (톤)	용도
충돌	9	4~8	남해,서해,항구	5~10	어선
침몰	9	4-8, 8-12	남해,서해,항구	1~2	어선, 기타선, 레저기구
침수	9	8-12, 4-8	남해,서해,항구	1~2, 5~10	어선
접촉	10	모두 비슷	남해, 항구	1000~5000	어선, 레저기구, 예인선, 화물선
기타	10	20-24, 4-8	남해, 항구	100~500	어선, 기타선
해양오염	3,4,7,8	8-12, 12-16	항구	100~500, 50~100	어선, 유조선, 화물선, 예인선
좌초	10	8-12, 4-8	남해, 서해	5~10	어선
추진축계손상	9	12-16, 8-12	서해, 남해	5~10	어선, 레저기구
안전사고	10	8~12	서해	5~10	어선
운항저해	10	12-16, 8-12	남해, 서해	1~2	어선, 레저기구
부유물감김	11,10,12	12-16, 8-12	남해	5~10, 3~5	어선
화재 · 폭발	1,4,8,11	12-16, 8-12	남해	5~10	어선
기관손상	9	12-16, 8-12	남해	5~10, 1~2	어선
전복	9	8~12	남해	1~2	어선
조타장치손상	모두 비슷	8-12, 12-16	남해	5~10	어선

(표 4) 사고의 유형 별 변수의 우선순위

주된 차이는 톤 수를 통해 알아볼 수 있었다. 접촉, 해양오염, 기타의 경우 대형 선박을 위주로 발생하였음을 알 수 있으며 아주 작은 소형선박들은 기관손상, 운항 저해, 전복, 침몰, 침수 등의 사고를 겪고 있음을 알 수 있다. 또한 사고가 일어나는 선박의 용도는 어선이 주를 이루며 이는 어선의 항해율이 가장 높으나 기타 선박에 비해 안전에 대한 철저한 대비가 이뤄지지 않고 있음을 유추할 수 있다. 월, 시간대의 경우 비슷한 개형을 보이나 약간씩 차이가 존재하였다. 최종적으로 이를 통해 선박의 크기 및 용도, 항해 시간 및 지역에 따라 사고의 유형이 변화한다는 사실을 파악할 수 있었다.

표에서 동일한 색으로 채워져 있는 유형은 유사한 변수를 가진다. 이는 각 변수마다의 상관관계가 존재하여 특정 변수에서 자주 나타나는 또 다른 변수가 존재함을 알 수 있다. 충돌, 침몰, 침수 사고는 남해, 서해, 항구 등 다양한 지역에서 보이는 현상이며 주로 9 월 4~12 시에 발생한다. 접촉, 기타, 해양 오염의 경우 톤 수의 범위 크고 용도가 다양하며 주로 항구에서 사고가 발생한다 이는 대형 선박의 경우 항해로가 겹치지 않아 항구로 진입 시 많은 사고가 발생함을 유추 가능하다. 좌초, 추진축계 손상, 안전 사고, 운항 저해의 경우 남해와 서해에서 가을에 주로 발생한다. 레저기구의 많은 비율이 해당하며 이는 가을철 레저스포츠를 즐기는 인원이 증가하며 생기는 현상임을 유추할 수 있다. 부유물 감김과 화재 폭발 사고는 주로 겨울에 나타나며 8~12 시에 발생한다. 기관 손상, 전복, 조타장치 손상은 8~16 시, 남해 소형 어선에서 많이 발생함을 알 수 있다. 이처럼 동일하게 분류 되어있는 사고의 유형은 모두 비슷한 원인과 종류, 피

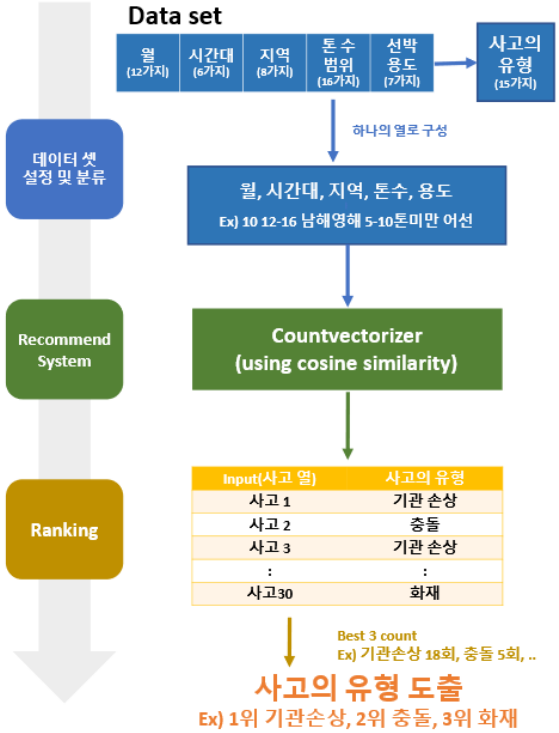
해 정도를 나타냄 또한 알 수 있다.

결론적으로 시간, 장소, 선박의 정보는 각각의 상관관계를 가지며 사고의 유형 결정에 영향을 주고 있다고 판단하였다. 따라서 전체 사고데이터 중 월, 시간대(통계용), 지역(통계용), 톤 수 범위(통계용), 선박 용도를 input 값으로 취하고 사고의 유형을 output 값으로 가지도록 데이터 셋을 구성하였다.

3.2 알고리즘의 구성

추천 알고리즘이란 데이터를 통해 사용자가 아직 소비하지 않은 아이템 중, 선호할 만한 아이템을 예측하는 알고리즘이다. 대표적인 추천 알고리즘으로는 contents based filtering, collaborative filtering, hybrid filtering 이 있으며 본 프로젝트에서는 contents based filtering 을 활용한다. 이는 각 사용자와 아이템에 대한 프로필 작성을 기반으로 하는 기법이다. 이는 user-based 와 item-based 로 구분할 수 있다. 사용자를 기반으로 할 경우 나와 비슷한 프로필의 다른 사용자의 선호를 추천한다. 아이템을 기반으로 할 경우는 특정 아이템을 선호했던 사용자에게 비슷한 아이템을 추천해준다. 하지만 데이터셋, 즉 프로필 작성이 어렵다는 단점을 가지고 있고, 평가가 주관적일 경우 객관성이 떨어진다는 문제를 가지고 있다.

하지만 본 논문에서 진행하고자 하는 프로젝트의 경우, 사고의 유무로만 판단을 하기 때문에 주관성이 개입되지 않고 기존의 사고데이터가 프로필처럼 구성되어 양이 많기 때문에 contents based 필터링을 사용하더라도 크게 무리가 없으므로 이를 채택하였다.



<그림 2> 유형 알고리즘 flowchart

사고의 데이터를 모두 하나의 열로 축약 후 countvectorizer 을 동일한 상황의 사고들을 뽑은 후, 해당 데이터의 사고 유형을 파악하여 "가장 비슷한 상황에서 벌어진 사고의 유형" 을 추출하였다. 이후 30 개의 가장 비슷한 사고들을 유사도가 높은 순서대로 나열 후 가장 높은 비율을 차지하는 사고의 유형을 최종 결과로 제공한다. (또한 출력된 사고의 개수에 따라 순위를 나누어 발생가능한 사고의 종류를 적게는 1 개, 많게는 3 개까지 제공하였다.)

3.3 결과 분석

사고(No.)	실제 유형	출력 횟수	정답 비율	사고(No.)	실제 유형	출력 횟수	정답 비율
1	기관 손상	25/30	0.83	11	전복	26/30	0.87
2	충돌	20/30	0.67	12	운항저해	17/30	0.57
3	기관손상	25/30	0.83	13	충돌	10/30	0.34
4	기관손상	21/30	0.70	14	기관손상	19/30	0.63
5	충돌	21/30	0.70	15	침수	13/30	0.43
6	안전사고	17/30	0.57	16	운항 저해	25/30	0.83
7	침몰	14/30	0.47	17	충돌	29/30	0.97
8	충돌	19/30	0.63	18	안전사고	26/30	0.87
9	기관손상	20/30	0.67	19	기관손상	29/30	0.97
10	화재 폭발	9/30	0.30	20	충돌	20/30	0.67

(표 5) 유형 알고리즘 결과 분석

20 회의 테스트 결과를 분석해본 결과 평균적으로 30 개의 사고 유형 중에서 20.25 회의 사고는 실제 사고 유형과 동일한 유형을 보였다. 이는 최종적으로 3 개의 일어날 수 있는 사고의 유형을 제공함에 있어 큰 어려움이 없을 것이라 판단하였다.

4. 결론

본 논문에서는 인공지능 기술 중 로지스틱 분류를 활용하여 일어날 수 있는 사고의 확률을 예측하고 추천 알고리즘을 통해 사고의 유형을 예측하는 시스템을 설계 및 구축하였다. 이는 추후 소형 선박에 제공될 자율 주행 및 e-navigation system 에 활용될 수 있을 것이라 기대되며 사고를 미리 예방하여 사고의 발생 확률을 낮추는 데 큰 도움을 줄 수 있을 것이다.

참고문헌

[1] 관계부처합동 보도자료(2019.01.08),「스마트 해상물류 체계구축전략(안)」,『과기관계장관회의』, 2019-01(2), 과학기술정보통신부

[2] 서정환· 김장원· 정동원,「최적 항로 탐색 지원을 위한 자율운항선박 모형 설계」,『하계공동학술대회』, 한국정보기술학회, (2019), p148-151

[4] 해양경찰청,「2019 년 해상조난사고 통계연보」, (2019), 11-153000-00046-10, p23

본 논문은 해양수산부 실무형 해상물류일자리지원사업의 지원을 통해 수행한 ICT 멘토링 프로젝트의 결과물입니다