ISSN: 3022-5388
JKAIA website: https://accesson.kr/jkaia
DOI: http://dx.doi.org/10.24225/jkaia.2023.1.2.1

시각적 특징과 머신 러닝으로 악성 URL 구분: HTTPS의 역할

Malicious URL Detection by Visual Characteristics with Machine Learning: Roles of HTTPS

Sung-Won HONG[1], Min-Soo KANG[2]

Received: October 26, 2023. Revised: November 16, 2023. Accepted: December 30, 2023

Abstract

In this paper, we present a new method for classifying malicious URLs to reduce cases of learning difficulties due to unfamiliar and difficult terms related to information protection. This study plans to extract only visually distinguishable features within the URL structure and compare them through map learning algorithms, and to compare the contribution values of the best map learning algorithm methods to extract features that have the most impact on classifying malicious URLs. As research data, Kaggle used data that classified 7,046 malicious URLs and 7.046 normal URLs. As a result of the study, among the three supervised learning algorithms used (Decision Tree, Support Vector Machine, and Logistic Regression), the Decision Tree algorithm showed the best performance with 83% accuracy, 83.1% F1-score and 83.6% Recall values. It was confirmed that the contribution value of https is the highest among whether to use https, sub domain, and prefix and suffix, which can be visually distinguished through the feature contribution of Decision Tree. Although it has been difficult to learn unfamiliar and difficult terms so far, this study will be able to provide an intuitive judgment method without explanation of the terms and prove its usefulness in the field of malicious URL detection.

Keywords : Decision Tree, Support Vector Machine, Linear Regression, URL, Visual feature

Major Classification Code : URL, Decision Tree ,Visual feature

1. Introduction

통계청의 데이터에 따르면, 2023년 현재 사용 중인 스마트폰 수는 약 5,500만대로 추정되며 해당 수치는 2019년 이후 지속적으로 증가하고 있다(KOSIS, 2023). 더불어, 대한민국 국민 수가 약 5,200만 명인 점을 고려하면 평균적으로 개인당 1대 이상의 스마트폰을 사용 중인 것으로 추정할 수 있다. 스마트폰의 발전은 현대 사회의 편의성을 증가시켰으나, 이러한 발전과 함께 사용자의 정보를 해킹하려는 방법 또한 다양해지고 있다.

스마트폰을 통해 사용자의 정보를 해킹하는 대표적인 방법 중 하나는 '보이스피싱'이다. 보이스피싱은 스마트폰을 통해 모바일 이체가 가능하다는 점을 악용한다. 특히 2012년 이후 검사를 사칭하여 돈과 정보를 탈취하는 사례가 지속적으로 증가하였으나, 국가 차원에서의 교육과 관심으로 2019년 이후로는 보이스피싱 발

1 First Author. Medical IT, Eulji University, Republic of Korea. Email: sungwon188@daum.net
2 Second Author. Professor, Medical IT, Eulji University, Republic of Korea. Email: mskang@eulji.ac.kr

생 건수가 감소하는 모습을 보이고 있다.

그러나 또다른 해킹 공격 방법으로 '스미싱' 이 등장하였다. 스미싱은 'SMS'와 'Phishing'의 합성어로 정상 URL과 유사한 특징을 가진 URL 형태면서 URL을 클릭할 시 정보를 해킹하는 악성 URL이 포함된 문자 메시지를 통해 공공기관, 택배 회사, 가족 등을 사칭하여 사용자를 속이고 악성 URL을 클릭하도록 유도하여 사용자의 정보를 해킹하고 돈을 탈취하는 방법이다(AhnLab, 2023). 현재 악성 URL과 정상 URL을 구분하는 연구도 진행되고 있으며 국가에서도 예방 교육을 진행하고 있음에도 불구하고, 스미싱 피해 사례는 2019년 이후 현재까지 지속적으로 증가하고 있다. 과학기술정보통신부에서 발표한 정보보호 실태조사에 따르면, 현재 정보보호 관련 용어가 생소하고 어려워 학습에 어려움을 겪는다고 한다(KOSIS, 2023). 이러한 문제 해결을 위해 직관적 판단이 가능한 분류 기준을 정하고자 하였다. 본 연구에서는 시각적 판단이 가능한 특징을 기반으로 지도학습 알고리즘을 적용하여 각 특징이 악성 URL 분류에 미치는 영향을 비교하였다.

2장에서는 악성 URL을 분류하기 위해 진행되었던 관련 연구 및 배경지식을 설명하였고, 3장에서는 연구에서 사용된 툴과 연구 과정에 대해 설명하였다. 4장에서는 연구를 통해 얻은 결과를 비교하였고, 연구 결과를 통해 5장에서 결론을 작성하였다.

2. Related Research and Background Knowledge

2.1. Related Research

한채림 등은 머신러닝 기반의 특징 선택 알고리즘을 사용해 악성코드를 분류하였고, 가중 유클리드 거리를 활용하여 사전처리를 진행한 후 난독화 요소를 제거하여 정확도를 개선하였다(Han et al., 2022). 학습 데이터로는 정상 URL 946개, 악성 URL 1,298개를 사용하여 진행하였으며, 89.17%의 정확도를 보여준 연구이다.

김영준 등은 기존에 제안된 URL 길이, 문자와 숫자 그리고 특수문자 개수, 문자열 포함 여부의 URL 특징에다가 URL Days, URL Words, URL Abmormal 등의 특징을 추가하여 머신러닝을 개발하였다(Kim & Lee, 2022). 실험에 사용된 데이터 세트는 정상과 악성 URL을 1:1 비율로 총 100,000개를 사용하여 진행하였고, 8:2의 비율로 학습 데이터와 테스트 세트로 분류하였다. Decision Tree 등 4개의 머신러닝 알고리즘을 적용시켜 진행하였고, 98.5%의 정확도를 보여준 연구이다.

정준영 등은 유해사이트의 HTML 코드를 수집하여 유해사이트의 특징을 분석하였고, 해당 특징을 이용하여 유해 사이트를 구분하였다(Jang et al., 2022). 유해사이트의 HTML 코드 특징으로는 도메인에 붙는 시퀀스 번호 유무, HTML 메타 데이터의 존재 유무, 하이퍼링크를 포함한 이미지, 타이틀에 사용되는 키워드가 있다는 것을 분석을 통하여 알아내었다. 이러한 4가지의 특징을 통해 743개의 유해사이트를 식별하였으며, 98.79%의 높은 정확도를 보여준 연구이다.

김종관 등은 악성 URL을 실시간으로 탐지하기 위한 검색량 기반 악성 URL 탐지 기법을 제안하였다(Kim et al., 2021). 학습 데이터는 정상 데이터 1,000,000건, 비정상 데이터 14,767개를 이용하였고, 머신러닝 알고리즘 (SVM, KNN, LR)을 적용시켜 탐지 기법을 개발하였다. 테스트 데이터로는 정상, 비정상 데이터 모두 200개로 구성하였고, 머신러닝 알고리즘 적용 결과 KNN의 평균 탐지 정확도가 88%로 가장 높다는 것을 보여준 연구이다.

강홍구 등은 URL의 어휘적 특징을 이용하여 머신러닝 알고리즘 모델을 생성하고, 모델들의 조합을 통해 URL의 악성 여부를 예측하는 시스템을 개발했다(Kang et al., 2020). 예측 시스템의 성능을 측정하기 위해 정상 URL 데이터 436,722개, 악성 URL 데이터 158,081개를 수집하였고, 테스트 데이터는 정상 URL 데이터 87,246개, 악성 URL 데이터 31,715개를 사용하였고, 단일 머신러닝 모델보다 다양한 머신러닝 모델이 예측 성능 향상에 더 유용함을 증명한 연구이다.

김보민 등은 악성 URL을 문자로 받았을 때의 행동을 설문조사를 통해 데이터를 수집하고, 데이터를 URL-based Rule, HTML-based Rule 등 10가지 기준으로 나누었다(Kim et al., 2020). 10가지 기준을 기반으로 사이트의 악성 확률을 반환해 주는 HypoChain 1,1,2 라이브러리를 고안했고, 2가지 중국 URL을 통해 테스트해 본 결과 0%와 33%로 10가지 기준을 통해 악성 URL의 확률을 보여준 연구이다.

2.2. Azure ML

본 연구에서 사용한 Azure ML(Machine learning)은 프로젝트 수명주기를 가속화하고 관리하기 위한 클라우드 서비스이다(Microsoft, 2023). Azure ML은 인공지능 (AI) 및 기계 학습 (ML) 서비스를 제공하여 데이터 분석 및 AI 모델을 구성하는데 도움을 준다.

2.3. URL Structure

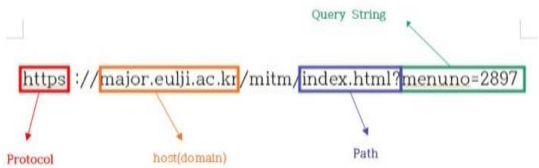

Figure 1: URL Structure

Figure 1과 같이 URL(Uniform resource locator) 구조는 프로토콜, host, port, path, query string로 구성되어 있다. 프로토콜은 상호 간 정의한 규칙으로 인터넷에서는 통신규약을 의미하며, HTTP와 HTTPS가 대표적이다. HTTP(HyperText Transfer Protocol)는 네트워크 통신을 작동하게 하는 기본 기술이다. HTTPS(HyperText Transfer Protocol Secure)는 HTTP의 확장 버전이다. HTTPS는 HTTP와 달리 암호화를 책임지는 SSL(Secure Socket Layer)인증서가 동작한다. SSL 인증서 기술은 암호화 방식으로 발신자와 수신자만 암호를 해독하여 정보를 주고받을 수 있는 기술이다. Table 1과 같이 보안을 중요시하는 금융 사이트 또는 공공기관 사이트들은 대부분 HTTPS를 이용하고 있으며, 이를 통해 진짜 은행 및 국가기관 사이트 인지, 가짜 은행 및 국가기관 사이트인지 분류할 수 있는 것이다.

Table 1 : URL of banks and public institutions

Name of the institution	URL
National police Agency	https://www.police.go.kr/index.do
Korea Customs Service	https://www.customs.go.kr/kcs/main.do
Board of Audit and Inspection	https://www.bai.go.kr/bai/
National Tax Service	https://www.nts.go.kr/
Shinhan Bank	https://www.shinhan.com
Woori Bank	https://www.wooribank.com/
Kookmin Bank	https://www.kbstar.com/

host는 접근하고자 하는 컴퓨터의 주소를 의미하며, 직접 IP주소를 사용 가능하다. port는 웹서버에서 자원을 접근하기 위해 사용하는 관문으로, 특정 서버 프로그램의 지정번호를 말한다. 마지막 query string은 URL에 추가적인 값을 담을 수 있는 공간으로, query string의 데이터를 기반으로 DB에 요청하거나 다른 동작을 요청하는 과정을 수행한다.

3. Research

3.1. Data Collection and Preprocessing

데이터 셋으로는 Kaggle에 업로드 되어있는 SANDUN ABEYSOORIYA의 악성 URL 특징 구분 데이터 셋을 사용하였다. 해당 데이터는 악성 URL, 정상 URL 14,092개(악성 URL 7,046개, 합법 URL 7,046개)를 각 URL 특징(URL길이, HTTPS존재 유무, 보조 도메인 존재 여부, 페이지 랭킹 순위 등)마다 정상 URL인지, 악성 URL인지 구분한 결과값을 0과 1로 입력해 놓은 데이터 셋이다. 지도학습 알고리즘을 적용하기 전에 데이터 세트를 생성하기 위한 과정이 이루어져야 한다. 먼저, 전처리로는 결측치, 이상치 데이터를 파악하여 수정하고, 총 29개의 칼럼 중 URL구조 중 시각적으로 보고 바로 판단할 수 있는 칼럼만 사용하였다.

URL의 여러 특징 칼럼 중에서 위와 같이 시각적으로만 판단할 수 있는 칼럼인 HTTPS, 다중 하위 도메인 보유 여부, port 번호, (-)이 포함된 접두사 및 접미사 추가 여부인 4개의 칼럼을 선정하였고 이를 시각적 특징으로 정의하였다. Azure에서는 Select column in Dataset 블록을 통해 사용되는 칼럼 설정을 진행하였다. 칼럼을 선정하고 난 후 데이터의 값을 정수형(int)이 아닌 문자 타입으로 변경하였고, 지도 학습에서 중요한 정답을 구분하는 Label 과정을 진행하였다. 본 연구에서는 URL이 악성 URL 인지, 정상 URL 인지 구분하는 작업을 진행하기에 학습 데이터의 자료 중 Result 값을 Label로 설정하였고 Azure ML에서는 Edit Meta data 블록을 통해 진행하였다.

3.2. Model Configuration

전처리를 수행한 데이터를 구분하는 Split Data 블록을 진행하였다. Split Data 과정에서는 데이터 셋을 훈련 데이터와, 평가 데이터 8:2 비율로 구분하였고, 훈련 데이터와 평가 데이터에 포함되는 데이터들은 랜덤으로 구분되었다. 이는 14,092개의 데이터 중 80%를 랜덤으로 추출하여 학습 데이터로 사용하고, 나머지 20%를 평가 데이터로 적용한 것을 의미한다. Data split 과정 이후, 훈련 데이터를 지도 학습 알고리즘 종류인 Two-class Decision Tree, Two-class Support Vector Machine, Two-class Logistic Regression 알고리즘에 학습을 진행하였다.

Two class Decision Tree는 질문을 던져서 맞고 틀리는 것에 따라 대상을 좁혀 나가는 방법으로, 분류 분석 방법으로 사용하였다. Two-class Support Vector Machine은 두 클래스로부터 최대한 멀리 떨어져 있는 결정 경계를 찾는 분류기로 특정 조건을 만족하는 동시에 클래스를 분류하는 것을 목표로 하는 방법으로, Decision Tree와 같이 분류 분석 방법으로 사용하였다. 마지막, Two-class Logistic Regression은 두 데이터 요인 간의 관계를 찾는 데이터 분석 기법으로, 위의 두 종류와 같이 분류 분석 방법으로 사용하였다.

각 학습 알고리즘을 학습 데이터를 통해 학습을 진행하였고, 진행 과정 및 결과물은 Train model 블록을 통해 확인할 수 있다. 학습을 완료하였으면, 예측한 결과 값이 정확한지, 부정확한지를 평가하는 단계가 진행된다. 이는 Score model 블록을 통해 확인할 수 있으며, 각 예측 결과와, 예측 수치를 확인할 수 있다. 마지막으로, 예측 결과를 통해, 각 특징이 얼마나 URL을 분류하는데 영향을 주는지에 대한 정보를 보여주는 Evaluate model 블록을 진행하는 과정을 거친다.

4. Research Results

4.1. Comparison of Algorithm Results

본 연구의 결과를 비교하기 위해 Azure에서 Evaluate Model 과정을 진행하였고, 평가 지표로는 F1-score, Recall, 정확도(Accuracy) 지표를 사용하였다.

F1-score, Recall, Accuracy 지표 결과를 비교하기 전에 각 지표에서 사용되는 Confusion Matrix에 대해 설명하고, 각 지표가 어떤 공식으로 계산되었고, 어떤 의미를 가지고 있는지 알아보았다.

3개의 지표는 Confusion Matrix를 이용하여 계산된 결과이기 때문에 Confusion Matrix의 의미도 이해할 필요가 있다. Confusion matrix는 교차검증에서 테스트 데이터의 예측과 정답 Label이 얼마나 잘 분류되었는지를 판단하기 위해 사용되는 지표로, Table2와 같이 TP, TN, FP, FN으로 구성되어 있다. TP는 True Positive의 줄임말로 테스트 데이터를 정답으로 예측하였는데, 실제 결과가 정답인 경우이고, TN은 True Negative의 줄임말로 테스트 데이터를 오답으로 예측하였는데, 실제 결과가 정답인 경우이다. FP는 False Positive의 줄임말로 테스트 데이터를 정답으로 예측하였는데, 결과는 정답이 아닌 경우이고, FN은 False Negative의 줄임말로 테스트 데이터를 오답으로 예측하였는데, 결과가 오답인 경우를 의미한다.

Table 2: Confusion Matrix

	Positive	Negative
Positive	TP	FN
Negative	FP	TN

$$F1-score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (1)$$

F1-Score는 보통 불균형 분류 문제에서 평가 척도로 사용되며 수식(1)로 나타낸다. 데이터가 불균형 상태에서 Accuracy로 성능을 평가하기엔 데이터 편향성이 너무 크게 나타나기에 올바르게 성능을 측정하기 힘들기 때문에, Sensitivity와 Precision을 이용하여 조화평균을 이루고 있는 F1-Score를 평가 척도로 사용한다.

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \quad (2)$$

Recall은 Sensitive라고도 불리며, True Positive Rate로 실제 Positive를 얼마나 잘 예측하는지를 나타내는 지표이며 수식 (2)로 나타낸다. Recall은 실제 정답을 예측하는 비율로 높을수록 좋은 척도라고 할 수 있다.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} \quad (3)$$

Accuracy는 정확도로 전체 예측한 것 중에 올바른 예측을 얼마나 했는지를 나타낸 지표이며 수식 (3)으로 나타낸다. Accuracy는 TP, TN, FP, FN 4가지 경우 중에서 테스트 데이터를 정답이라고 예측한 경우를 구하는 방법이다. 따라서 Accuracy는 Recall과 마찬가지로 높을수록 좋은 척도라고 할 수 있다.

Table 3: Results of F1-score, Recall, Accuracy

	F1-score	Recall	Accuracy
DT	83.1	83.6	83.0
SVM	81.6	80.4	81.8
LR	79.7	77.4	80.3

세 개의 지도학습 알고리즘(Decision Tree, Support Vector machine, Linear Regression) 모두 지표가 80%에 근사한 값을 갖는 준수한 결과를 나타냈다.

ROC커브는 분류 분석 모형의 평가를 쉽게 비교할 수 있도록 시각화한 그래프이다. ROC커브의 x축은 False positive Rate이고, y축은 True positive Rate로 구성되어 있다. ROC 커브의 아래 면적의 값이 1에 가까울수록 모형의 성능이 우수한 것으로 Two-Class Decision Tree의 ROC 커브 모양처럼 생긴 모양이 좋은 모양이다. 따라서 ROC 커브에서도 Two-class Decision Tree가 가장 좋은 학습 방법임을 알 수 있다.

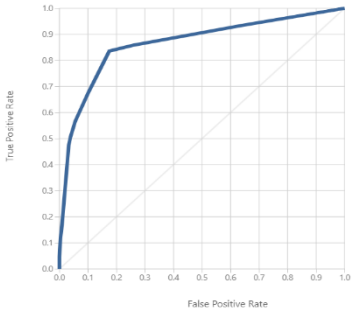

Figure 1: Roc curve of Two-class Decision Tree

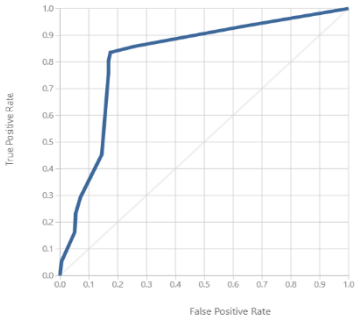

Figure 2: Roc curve of Two-class Support Vector Machine

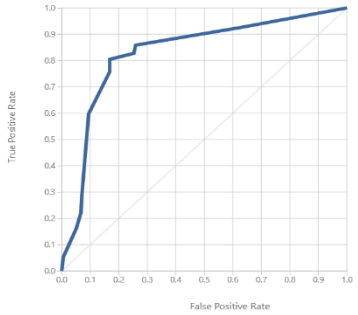

Figure 3: Roc curve of Two-class Logistic Regression

Confusion Matrix를 통해 Decision Tree, Support Vector Machine, Linear Regression 결과 값을 분석해 보았다. F1-score의 결과값이 83.1 %, 81.6 %, 79.7% 로 Decision Tree가 가장 높은 것을 알 수 있고, 이를 통해 Decision Tree가 불균형하게 분류되어 있는 것이 아니라, 균형 있게 분류가 잘 되어있음을 알 수 있다.

또한, Recall과 Accuracy를 비교해 본 결과 Decision Tree의 결과가 83.6%와 83.0%로 가장 높은 것을 알 수 있다. 따라서 세 가지 지도학습 알고리즘 중 Decision Tree 학습 방법이 분류도 잘 되어있고, 정확도와 재현성이 높아 좋은 척도라는 결과를 알 수 있다.

4.2. Analysis of Two-class Decision Tree Results

세 개의 지도학습 알고리즘 중에서 가장 결과가 좋은 Two-class Decision Tree를 분석해 보았다. 분석해 본 결과 figure 5를 보면 HTTPS의 Feature Score가 0.26으로 Two-class Decision Tree가 악성 URL을 분류하는데 HTTPS의 존재 유무가 가장 많은 영향을 준다는 것을 알 수 있었다. 또 비표준 포트 사용의 Feature Score가 0인 것을 알 수 있었고, 수치가 0이기에 figure 5에서는 비표준 포트 사용의 값을 제외하였다.

지도학습을 식으로 작성한다면 Label과 가중치의 값과, Feature, bias값을 기반으로 $y = w \times x + b$ 식으로 작성이 가능하다. 위의 Decision Tree의 Feature 결과 값을 기반으로 작성하면 *악성URL 결과* $= 0.26 \times$ *https의 유무* 가URL의 시각적인 특징을 기반으로 지도학습 알고리즘을 적용한 분류의 공식임을 알 수 있다. bias는 Feature Score 값에 따로 적용된 값이 없기에 bias의 값은 추가하지 않았다.

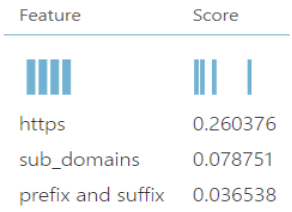

Figure 4: Two-class Decision Tree Feature Contribution

5. Conclusion

본 논문에서는 시각적 분류만을 통해 악성 URL을 식별할 수 있도록, 시각적 특징을 기반으로 Decision Tree, Support Vector Machine (SVM), Linear Regression 등의 지도학습 알고리즘을 활용하여 분석하였다.

분석 결과, Decision Tree가 가장 높은 정확도, F1-score, Recall 값을 보였다. 이러한 알고리즘의 학습 방법을 분석한 결과, HTTPS의 유무가 악성 URL을 구분하는 데 가장 큰 영향을 주는 것으로 확인하였다.

본 연구 결과를 바탕으로 복잡한 용어에 대한 이해가 없더라도 HTTPS의 존재 유무만으로도 악성 URL을 판단할 수 있다는 전제가 타당함을 알 수 있다. 따라서 현재 사용되는 교육 자료의 복잡한 용어 대신 사용된다면 용어를 이해하는 어려움이 줄어들 뿐만 아니라, 스미싱 피해도 자연스럽게 줄어드는 좋은 효과를 얻을 수 있을 것으로 기대된다.

향후에도 복잡한 용어를 이해할 필요 없이, 시각적인 특징만으로 악성 URL을 구분할 수 있도록 악성 URL에 대한 지속적인 연구가 이뤄져야한다.

References

AhnLab. (2023, July 18). Retrieved November 5, 2023, from https://www.ahnlab.com/ko/contents/content-center/33769

Han, C. R., Yun, S. H., Han, M. J., & Lee, I. G. (2022). Machine Learning-Based Malicious URL Detection Technique. *Journal of the Korea Institute of Information Security & Cryptology*, *32*(3), 555-564.

Jang, J. Y., Lim, K. D., & Lee, S. J. (2022). An Harmful site collection system using Characteristic of HTML and URL. *Journal of Digital Forensics, 16*(1), 54-63.

Kang, H. K., Shin, S. S., Kim, D. Y., & Park, S. T. (2020). Design and Implementation of Malicious URL Prediction System based on Multiple Machine Learning Algorithms. *Journal of Korea Multimedia Society*, *23*(11), 1396-1405.

Kim, B. M., Han, Y. W., Kim, G. Y., Kim, Y. B., & Kim, H. J. (2020). Development of Rule-Based Malicious URL Detection Library Considering User Experiences*. *Journal of the Korea Institute of Information Security & Cryptology*, *30*(3), 481-491.

Kim, J. K., Jang, M. H., Lim, S. N., & Kim, M. S. (2021). A Study on the Detection Method of Malicious URLs based on the Internet Search Engines using the Machine Learning. *The transactions of The Korean Institute of Electrical Engineers*, *70*(1), 114-120, 10.5370/KIEE.2021.70.1.114

Kim, Y. J., & Lee, J. W. (2022). Development of a Malicious URL Machine Learning Detection Model Reflecting the Main Feature of URLs. *Journal of the Korea Institute of Information and Communication Engineering*, *26*(12), 1786-1793.

KOSIS. (2023, March 7). Retrieved November 3, 2023, from https://kostat.go.kr/ansk/

KOSIS. (2023, August 25). Retrieved November 7, 2023, from https://kosis.kr/index/index.do

Microsoft. (2023, December 6). Retrieved December 10, 2023, from https://learn.microsoft.com/ko-kr/azure/machine-learning/overview-what-is-azure-machine-learning?view=azureml-api-2