ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 47, No. 10, pp. 906-910, 2020. 10
https://doi.org/10.5626/JOK.2020.47.10.906

개인정보의 비식별화에 따른 기계학습의 예측 정확도 분석 연구

(A Study on the Prediction Accuracy of Machine Learning using De-Identified Personal Information)

정 홍 주 † (Hongju Jung)　이 나 영 †† (Nayoung Lee)　설 수 진 †† (Soo-jin Seol)　한 경 석 ††† (Kyeong-Seok Han)

요 약 개인정보 보호 및 개인정보 보호법 개정에 따른 개인정보 비식별화 관련 사항이 대두되고 있다. 또한 4차 산업혁명의 원동력으로 인공 지능과 기계학습의 활용이 증대되고 있다. 본 논문에서는 k-익명성(k=2)을 적용한 비식별화 개인정보를 활용하여 기계학습의 의사결정나무 알고리즘으로 예측 정확도를 실험적으로 검증한다. 그리고 입력 데이터의 예측 결과를 비교하여 기계학습에서 비식별화 개인정보를 활용 시 제한 사항을 알아보고자 한다. 개인정보보호법 개정안에 따라 기계학습에 비식별화 개인정보를 사용할 경우, 개인정보 비식별화 수준과 분석 알고리즘을 고려하여 활용해야 한다는 것을 제안한다.

키워드: k-익명성, 기계학습, 비식별화, 의사결정나무

Abstract The de-identification of personal information is emerging due to the revision of the Personal Information Protection and Personal Information Protection Act. In addition, the use of artificial intelligence and machine learning is becoming a driving force in the Fourth Industrial Revolution. In this paper, we experimentally verify the predictive accuracy of a machine learning decision tree algorithm using de-identified personal information by applying k-anonymity (k=2). The prediction results of the input data are compared to determine the limitations of using de-identified personal information in machine learning. According to the amendment of the Personal Information Protection Act, we propose that when using de-identified personal information in machine learning, the level of personal information de-identification and the analysis algorithm should be considered.

Keywords: k-anonymity, machine learning, de-identification, decision tree

† 학생회원 : 숭실대학교 대학원 IT정책경영학과 학생
Hongjujung@outlook.com
†† 정 회 원 : 숭실대학교 대학원 IT정책경영학과 학생
best2409@naver.com
vip7777vip@naver.com
††† 정 회 원 : 숭실대학교 경영학부 교수(Soongsil Univ.)
kshan@ssu.ac.kr
(Corresponding author임)

논문접수 : 2020년 2월 5일
(Received 5 February 2020)
논문수정 : 2020년 8월 21일
(Revised 21 August 2020)
심사완료 : 2020년 8월 27일
(Accepted 27 August 2020)

정보과학회논문지 제47권 제10호(2020. 10)

1. 서 론

4차 산업 혁명의 일환으로 인공지능, 데이터 등에 대한 관심이 증대되고 있으며, 산업 분야별로 활용이 많아지고 있다. 인공지능은 딥러닝, 기계학습 등 다양한 분야로 발전하고 이를 활용하여 제품이나 IT 서비스에 이미 적용되고 있다. 또한 21세기 원유라고 불리는 데이터에 대한 관심도 증가하고 있다.

반면에 개인정보가 포함된 데이터에 대해서는 개인정보 보호, 프라이버시 보호 등에 따라 이용하기 위해서는 정보주체의 동의가 있어야 하므로 데이터 유통이 제한되어 있어 데이터 분석 및 활용 측면에서 제약이 존재하고 있다. 하지만 개인정보 보호법 일부개정법률안이 가결됨에 따라 가명 정보의 경우 일부 목적에 한해 정보주체의 동의 없이 처리할 수도 있으며[1], 추가로 개인정보 비식별 조치 가이드라인[2] 활용이 가능하다.

하지만 원본 데이터를 변환, 삭제하여 생성되는 비식별화 개인정보를 통해 기계학습, 분석 등에 사용하게 되면, 비식별화 정도, 적용 알고리즘 등에 따라 상이할 수는 있지만 잘못된 학습이 될 수 있는 가능성이 존재한다.

본 논문에서는 비식별화 개인정보 데이터에 대한 기계학습의 예측 정확도를 분석하기 위해 개인정보 비식별 조치 가이드라인[2]의 프라이버시 모델에서 k-익명성(k=2)를 적용한 비식별화 데이터와 원본 데이터를 기계학습의 의사결정나무(Decision Tree) 알고리즘으로 학습하여 예측 정확도를 비교, 분석해보고자 한다.

논문의 구성으로 2장에서는 개인정보 보호법, 의사결정나무(Decision Tree)와 같은 배경 지식과 관련 연구를 설명하고 3장에서는 원본, 비식별화 데이터와 기계학습 모델을 제시하고, 이에 대한 실험 방법을 설명한다. 4장에서는 원본 데이터와 비식별화 데이터에 대한 기계학습 예측 결과를 비교 분석하고, 5장에서 결론을 도출한다.

2. 배경 지식 및 관련 연구

2.1 개인정보 및 개인정보 보호

개인정보 보호법[3]에서 '개인정보'는 살아 있는 개인에 관한 정보로서 성명, 주민등록번호 및 영상 등을 통하여 개인을 알아볼 수 있는 정보(해당 정보만으로는 특정 개인을 알아볼 수 없더라도 다른 정보와 쉽게 결합하여 알아볼 수 있는 것을 포함한다)로 정의하고 있다. 그리고 개인정보처리자는 정보주체의 동의를 받은 경우, 법률에 특별한 규정이 있거나 법령상 의무를 준수하기 위하여 불가피한 경우 등에 해당하는 경우에는 개인정보를 수집할 수 있고, 그 수집 목적의 범위에서 이용할 수 있으며, 정보주체의 동의를 받은 경우 등에 해당되는 경우에는 정보주체의 개인정보를 제3자에게 제공할 수 있도록 명시하고 있다[3]. 또한 개인정보 보호법 일부개정법률안[1]에 따라 통과된 개인정보 보호법[3]에서는 개인정보처리자는 당초 수집 목적과 합리적으로 관련된 범위 내에서 정보주체에게 불이익이 발생하는지 여부, 암호화 등 안전성 확보에 필요한 조치를 하였는지 여부 등을 고려하여 대통령령이 정하는 바에 따라 정보주체의 동의 없이 개인정보를 이용 또는 제공할 수 있도록 명시하고 있다.

가명 정보에 대한 내용이 추가된 개인정보 보호법[3]에서 '가명처리'는 개인정보의 일부를 삭제하거나 일부 또는 전부를 대체하는 등의 방법으로 추가 정보가 없이는 특정 개인을 알아볼 수 없도록 처리하는 것을 말하며, '가명정보'는 원래의 상태로 복원하기 위한 추가 정보의 사용·결합 없이는 특정 개인을 알아볼 수 없는 정보로 정의하고 있으며, 개인정보처리자는 통계작성, 과학적 연구, 공익적 기록보존 등을 위하여 정보주체의 동의 없이 가명 정보를 처리할 수 있으며 개인정보처리자는 가명정보를 제3자에게 제공하는 경우에는 특정 개인을 알아보기 위하여 사용될 수 있는 정보를 포함하지 않도록 명시하고 있다.

GDPR에서 '가명화'(pseudonymization)는 '추가적인 정보'(additional information)의 사용 없이 더 이상 특정 정보주체를 식별할 수 없는 방식으로 수행된 정보로 정의하고 있으며, 가명화를 거친 개인정보가 추가적인 정보의 사용에 의해 특정 개인의 속성으로 인정되는 경우는 개인정보로 간주하고 있다[4]. 개인정보를 외부에 제공하게 될 경우에는 데이터 자체의 특성과 상황별 개인정보 처리자를 고려하여 비식별화 수준을 결정해야 한다[5].

2.2 개인정보 비식별화 모델 및 기법

개인정보를 비식별 조치하여 이용 또는 제공하려는 사업자 등이 준수하여야 할 조치 기준을 제시한 개인정보 비식별화 조치 가이드 라인[2]에서는 사전검토, 비식별 조치, 적정성 평가, 사후 관리 단계 순서로 조치사항을 제공하고 있다.

적정성 평가 단계에서 k-익명성 모델을 활용하여 비식별 조치 수준의 적정성을 평가하게 된다. k-익명성은 특정인임을 추론할 수 있는 있는지 여부를 검토하여 일정 확률수준 이상 비식별되도록 하는 모델이며 동일한 값을 가진 레코드를 k개 이상으로 하여 특정 개인을 식별할 확률은 $1/k$이 되게 한다. 개인정보 비식별화를 위한 일반적 처리 기법으로는 가명처리, 총계처리, 데이터 삭제, 데이터 범주화, 데이터 마스킹하는 방법이 있다[2].

그리고 개별 질의에 대해서 응답에 잡음을 삽입하고, 통계 데이터를 배포할 경우 전체 데이터에 일부 잡음을 삽입하여 배포함으로서 개인정보를 보호하는 차분 프라이버시 모델 등에 대한 연구가 진행되고 있다[6].

2.3 기계학습의 의사결정나무

의사결정나무(Decision tree) 알고리즘은 의사결정규칙(Decision rule)을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류(Classification)하여 예측(Prediction)을 수행하는 분석방법으로 분석과정이 나무구조에 의해서 표현되기 때문에 판별 분석(Discriminant Analysis), 회귀 분석(Regression Analysis), 신경망(Neural Networks) 등과 같은 방법들에 비해 연구자가 분석과정을 쉽게 이해하고 설명할 수 있다는 장점을 가지고 있다[7]. 의사결정나무는 분류 또는 예측을 목적으로 하는 어떤 경우에도 사용될 수 있으나 분석의 정확도보다는 분석과정의 설명이 필요한 경우에 더 유용하게 사용된다[7].

Boosted 의사결정나무(Decision tree)는 두 번째 나무가 첫 번째 나무의 오류를 수정하고, 세 번째 나무가 첫 번째 및 두 번째 나무의 오류를 수정하는 등의 앙상

블 학습 방법이며 예측은 전체 트리의 앙상블을 기반으로 하여 예측을 수행한다. 일반적으로 적절히 구성된 경우 Boosted 의사결정나무(Decision tree)는 다양한 기계학습 작업에서 최상의 성능을 얻을 수 있는 가장 쉬운 방법이다[8].

2.4 관련 연구

비식별화 데이터가 기계학습에 미치는 영향에 관한 연구는 [9-11]이 있다. [9]에서는 Naïve Bayes 분류자를 통해 k 익명성이 적용된 데이터의 분류 정확도를 다양한 k 값에 대해 측정하였으며, 분류 정확도는 k 값이 5일 때 95%에서 k 값이 50일 때 90%로 나타내고 있다. [10]에서는 데이터에 k-익명성(k=2)를 적용한 후 다양한 기계학습 모델을 적용하여 예측율을 측정하였다. [11]에서는 로지스틱 회귀를 사용하여 원본데이터와 비식별데이터의 기계학습 예측도를 k-익명성(k=2)에 대해서는 원본대비 90%이상의 예측율을 나타내며, 큰 k 값과 ℓ 값에 대해서도 측정하였다.

기존 연구에서는 비식별 데이터에 대한 기계학습을 수행하지 않거나 비식별 데이터를 테스트 데이터로 사용하지 않아, 본 논문에서는 k-익명성(k=2)에 대해서 학습 및 테스트 데이터에 대한 원본 데이터(식별화), 비식별화 데이터를 통해 기계학습의 의사결정나무를 이용하여 식별화 및 비식별화 데이터에 대한 예측도와 범위 데이터에 대한 예측 정확도를 측정하고 비교한다.

3. 실험 및 평가 방법

3.1 실험 데이터

학습 및 테스트를 위한 데이터는 Kaggle의 Titanic 데이터의 Training set과 Test set을 이용한다[12]. 데이터는 탑승 승객에 대한 정보를 가지고 있으며(Name, Age, Sex, Pclass, SibSp, Parch 등), Training set은 학습을 위해 Survived 열로 생존여부를 제공하고 있고 Training set과 Test set의 데이터는 서로 상이하며, Age 열의 값 중에서 비어 있는 값은 사전에 제외하고 정수형으로 데이터 형식을 변환한다.

범위형 데이터에 대한 예측 결과를 비교하기 위해 나이에 대한 그룹 특성(0~20, 20~40 등)을 가진 AgeGroup 열을 추가하여 Training set과 Test set 원본 데이터(식별화 데이터)를 구성했다.

ARX 비식별화 도구[13]를 통해 Training set과 Test set 데이터의 비식별화를 적용한다. Name 열은 식별자이며, Age, Sex 열은 개인정보로 판단해서 k-익명성 모델로 k=2를 적용하여 Age 열은 범위 형태로, Sex 열은 범주 형태로 비식별화한다. 교차비교를 위해 Training set과 Test set 데이터를 동일한 방식으로 비식별화한다. ARX 비식별화 도구[13]에서 반환되는 최적의 비식별화 정도는 Training set과 Test set 데이터가 상이하게 나타난다. Age 열은 범위 형태로 "[0, 20[", "[20, 40[" 등과 같은 결과를 반환하므로 Age는 범위의 중간 값으로 변환하고 원본데이터와 동일한 구성으로 AgeGroup 열을 추가하여 Training set과 Test set 비식별화 데이터를 구성했다.

3.2 기계학습 모델

기계학습 모델은 Microsoft Azure Machine Learning Studio[14]를 이용하여 구축하였다. 입력데이터의 열 중에서 Survived, Pclass, Sex, Age, SibSP 열을 선택하고 8:2로 학습데이터를 분리하여 Two-class boosted Decision tree를 이용하여 학습을 수행하였으며 Two-class boosted Decision tree의 속성 중에서 Trainer mode는 Single Parameter로 설정하였다. 기계 학습의 결과로서 Scored Labels과 Scored Probabilities를 통해 Survived를 예측하였다. 기계학습 모델을 웹 서비스로 게시하고 해당 웹 서비스를 Excel에서 호출해 대량의 데이터에 대한 예측 결과를 도출하여 최종 측정 결과를 비교하였다.

3.3 모델 평가 방법

학습에 사용한 Training set 데이터에 대한 비식별화 데이터와 원본 데이터(식별화 데이터)의 예측 결과를 교차 비교하기 위해 학습 데이터만 다른 동일한 기계학습 모델을 구성하여 평가하였다.

평가를 위한 테스트 데이터는 Test set(식별화, 식별화)로 각각 기계학습 모델에 대한 예측 결과를 도출하여 4가지 형태로 생존(Survived=1)의 결과를 비교하였다.

그리고 도출된 예측 결과에 대한 Age 범위인 AgeGroup에 대한 정확도를 비교하여 평가하였다.

4. 실험 결과

그림 1은 기계학습 학습 데이터 유형별로 Test set의 식별화 데이터 대비 비식별화 데이터에 대한 생존(Survived =1) 예측 결과를 나타낸다. 원본 데이터(식별화)로 학습된 모델의 경우 Test set의 원본 데이터(식별화) 대비 비식별화 데이터는 60건(43%) 차이를 나타내고 있다. 반면에 비식별화 데이터로 학습된 모델의 경우는 19건 초과된 차이를 나타내고 있다.

Age 범위를 통해 더 구체적으로 결과를 확인해보면 그림 2는 원본 데이터(식별화)로 학습된 모델에서 AgeGroup 별 Test set의 원본 데이터(식별화) 대비 비식별화 데이터의 예측 정확도와 패턴을 확인할 수 있다. 전체적인 패턴은 동일한 모습을 나타내고 있다.

그림 3은 비식별화로 학습된 모델의 예측 정확도와 패턴을 확인할 수 있다. 전체적인 패턴은 유사하나 0~20 구간은 다른 차이가 나타내고 있다.

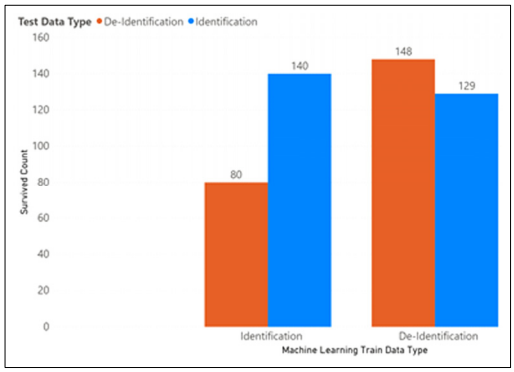

그림 1 원본 데이터(식별화) vs. 비식별화 테스트 데이터 예측 정확도

Fig. 1 Identification vs. De-Identification Test Data Prediction Accuracy

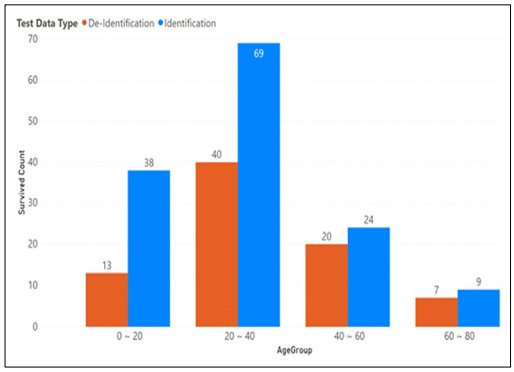

그림 2 원본 데이터(식별화)로 학습된 기계학습에서 AgeGroup별 예측 정확도

Fig. 2 Prediction Accuracy by AgeGroup on Machine Learning trained with Identification Data

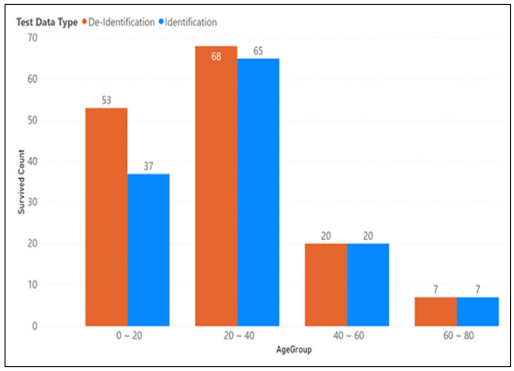

그림 3 비식별화 데이터로 학습된 기계학습에서 AgeGroup별 예측 정확도

Fig. 3 Prediction Accuracy by AgeGroup on Machine Learning trained with De-Identified Data

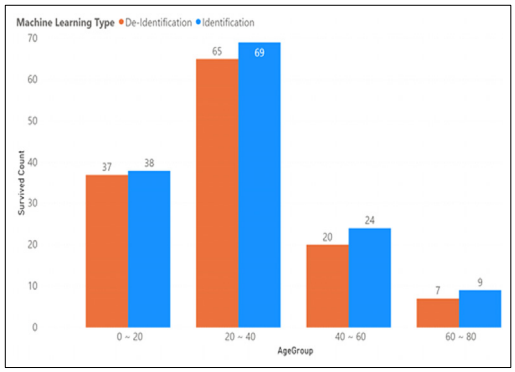

그림 4 원본 테스트 데이터(식별화)를 통한 기계학습에서 AgeGroup별 예측 정확도

Fig. 4 Prediction Accuracy by AgeGroup on Machine Learning using Identification Test Data

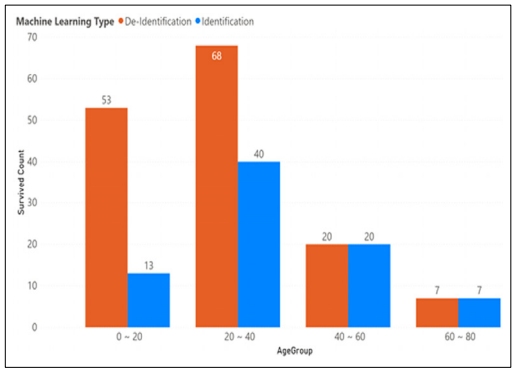

그림 5 비식별 테스트 데이터(식별화)를 통한 기계학습에서 AgeGroup별 예측 정확도

Fig. 5 Prediction Accuracy by AgeGroup on Machine Learning using De-Identified Test Data

그림 4와 그림 5의 경우 Test set의 비식별화, 식별화 데이터를 사용하여 학습 모델별로 예측 정확도와 패턴을 나타내고 있다. 그림 4의 경우 전체적인 패턴은 동일하게 나타내고 있다. 그림 5의 결과는 그림 4와는 달리 0~20, 20~40의 경우 다른 패턴을 나타내고 있다.

전체적으로 20~40이 가장 높으며 점차 감소하는 패턴을 공통적으로 보여주고 있다. 그림 2와 그림 4의 결과를 통해 원본 데이터(식별화)를 사용하여 학습하고 예측한 경우 범위별 패턴과 유사하다는 것을 알 수 있으며, 비식별화 데이터를 활용하여 학습하고 예측할 경우 잘못된 학습을 할 가능성이 있어 제한적이라고 판단할 수 있다.

5. 결 론

k-익명성(k=2)을 적용한 비식별화 개인정보를 활용하여 기계학습의 의사결정나무 알고리즘으로 예측 결과를

확인하였다. 기계학습에서 비식별화 데이터를 학습 데이터로 이용하거나 테스트 데이터 입력으로 활용할 경우와 비식별화하지 않은 원본 데이터를 활용하는 경우를 비교해본 결과, 예측 정확도와 범위별 데이터 패턴에서 차이가 나타났다. 이를 통해 비식별화 데이터를 이용할 경우 잘못된 결과를 도출될 가능성이 존재한다는 것을 확인할 수 있다.

개인정보를 기계학습에 활용할 때 분석 알고리즘, 분석 관점 등 상황과 데이터 특성을 고려하여 개인정보 비식별화 수준을 결정해야 한다. 본 연구에서 사용한 데이터 집합은 상황에 따라 Age 열에 대해 비식별화를 적용하지 않을 수 있으며, 그런 경우 예측 결과는 더 정확해지게 된다.

향후 연구 과제로서, 개인정보를 가명처리 하기 전에 기계학습 분석 알고리즘과 데이터 특성에 따라 어떤 개인정보를 어느 정도의 비식별화 수준을 적용할지 연구할 필요가 있다.

References

[1] BILL INFORMATION, Jan. 2020, Partial Amendment of the Personal Information Protection Act (Alternative), [Online]. Available: http://likms.assembly.go.kr/bill/billDetail.do?billId=PRC_L1Y9A1T1D2G5C1D8U1C3A3N5L7A6C1

[2] Guidelines for De-identifying Personal Information, Jun. 2016.

[3] Personal Information Protection Act [Enforcement Aug. 2020]

[4] General Data Protection Regulation (GDPR) Article 4 (5).

[5] J.-S. Kim, "Research on the Use of Pseudonym Data -Focusing on Technical Processing Methods and Corporate Utilization Directions-," *Journal of The Korea Institute of Information Security & Cryptology*, Vol. 30, No. 2, pp. 253-261, 2020. (in Korean)

[6] Korea Internet & Security Agency, "Systematization and establishment of privacy protection model to enhance personal information utilization," Apr. 2019.

[7] Jonghoo Choi and Doosung Seo, "Application of Data Mining Decision Trees," *Statistical Analysis of Statistics Korea*, Vol. 4, No. 1, pp. 61-83, 1999. (in Korean)

[8] Microsoft, May 2019, Two-Class Boosted Decision Tree, [Online]. Available: https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/two-class-boosted-decision-tree

[9] J. Paranthaman and T. Aruldoss Albert Vicroire, "Performance Evaluation of K-Anonymized Data," *Global Journal of Computer Science and Technology*, Vol. 13, 2013.

[10] H. Wimmer and L. Powell, "A Comparison of the Effects of K-anonymity on Machine Learning Algorithms," *International Journal of Advanced Computer Science and Applications*, Vol. 5, 2014.

[11] Min Kyoung Sung, Kang Jiseong and Jeong Seung Yu, "Analysis of Anonymized Data on Machine Learning," *KIISE Conference*, Vol. 2019, No. 6, pp. 106-108, 2019. (in Korean)

[12] Titanic Data, [Online]. Available: https://www.kaggle.com/c/titanic/data

[13] ARX Data Anonymization Tool, [Online]. Available: https://arx.deidentifier.org/

[14] Microsoft Azure Machine Learning Studio, [Online]. Available: https://studio.azureml.net/

정 홍 주

2003년 8월 숭실대학교 정보과학대학원 졸업(석사). 2019년~현재 숭실대학교 일반대학원 IT정책경영학과 박사과정. Microsoft MVP로 클라우드와 데이터 분석 컨설팅을 수행하고 있으며, 관심분야는 AI, 데이터분석 및 정보보안

이 나 영

2007년 8월 성균관 대학교 정보통신 대학원 졸업(석사). 2018년~현재 숭실대학교 일반대학원 IT정책경영학과 박사과정 현재 AWS 공인 강사. 관심분야는 클라우드 컴퓨팅, DevOps 분야

설 수 진

1998년 2월 경원대학교 동아 미술학과 학사 졸업. 2014년 2월 연세대학교 행정대학원 사회복지학과 졸업(사회복지학 석사). 2019년~현재 숭실대학교 일반 대학원 IT정책경영학과 박사과정. 2011년~현재 베스티안재단 사회복지사업본부 대표 1996년 제40회 미스코리아 선발대회에서 “선”으로 선발된 이후 방송 MC 및 드라마, 영화 등에서 활동

한 경 석

1979년 2월 서울대학교 문학사 졸업. 1983년 8월 서울대학교 경영학과 졸업(경영학 석사). 1989년 8월 미국 퍼듀대 MIS 박사 학위취득, 미국 휴스턴 대학교 조교수 역임. 1999년 The Wharton School 방문교수. 1993년 3월~현재 숭실대학교 경영학부 교수로 재직중. 관심분야는 경영정보 시스템(Technical MIS), Digital Economy, AI, Machine Learning, ERP, C++, Python, Java, 회계정보시스템, e-Business, 전자상거래, 중소기업정보화, 기업컨설팅, 정책 연구 등