

효율적인 병원보건관리를 위한 태아건강분류 모델

전제란

(의) 효성병원 경영관리원장

Design of Fetal Health Classification Model for Hospital Operation Management

Je-Ran Chun

Director of Hospital Management, Hyosung General Hospital

요 약 본 연구에서는 병원에서 실질적인 태아분만 시스템에 관리를 위한 태아건강분류모델을 설계하는 것을 목적으로 한다. 출산 중 사망자 수는 2017년을 기준으로 295,000명인 산모 사망률과 유사하다. 이러한 사망의 94%는 환경에 의해 발생하므로 대부분 예방할 수 있다. 따라서 본 논문에서는 랜덤 포레스트(Random Forest)를 이용하여 Cardiotocograms(CTG) 검사에서 추출한 2개의 데이터(태아의 심박수, 태아의 움직임, 자궁 수축 등)로 태아의 건강을 예측하는 모델을 제안하였다. 본 연구에서 제안된 모델은 태아분만 보건운영 시스템을 안정적으로 관리하기 위해 태아분만에 대한 데이터의 분포가 불균형한 이상 데이터를 갖는 항목을 찾아 표준편차의 상한 및 하한의 임계값을 설정하여 이상값을 제거하여 정확도를 높였다. 또한 태아의 건강상태를 나타내는 클래스의 비율이 불균형하므로, 데이터 리샘플링을 이용하여 소수의 클래스를 복제하여 클래스의 균형을 맞추었다. 그 결과 정확도가 4~5% 향상되어 97.75%로 나타났다. 이에 예측 모델을 통해 발생 할 수 있는 태아의 사망과 병을 사전에 정확히 예측하여 우선적으로 관리함으로써 효율적인 태아 보건운영과 태아 사망 및 병 예방에 기여할 수 있을 것이라고 기대한다.

주제어 : 태아건강, 분류모델, 랜덤 포레스트, 딥러닝, 리샘플링, 병원경영

Abstract The purpose of this study was to propose a model which is suitable for the actual delivery system by designing a fetal delivery hospital operation management and fetal health classification model. The number of deaths during childbirth is similar to the number of maternal mortality rate of 295,000 as of 2017. Among those numbers, 94% of deaths are preventable in most cases. Therefore, in this paper, we proposed a model that predicts the health condition of the fetus using data like heart rate of fetuses, fetal movements, uterine contractions, etc. that are extracted from the Cardiotocograms(CTG) test using a random forest. If the redundancy of the data is unbalanced, This proposed model guarantees a stable management of the fetal delivery health management system. To secure the accuracy of the fetal delivery health management system, we remove the outlier which embedded in the system, by setting thresholds for the upper and lower standard deviations. In addition, as the proportion of the sequence class uses the health status of fetus, a small number of classes were replicated by data-resampling to balance the classes. We had the 4~5% improvement and as the result we reached the accuracy of 97.75%. It is expected that the developed model will contribute to prevent death and effective fetal health management, also disease prevention by predicting and managing the fetus' deaths and diseases accurately in advance.

Key Words : Fetal health, Classification model, random forest, deep learning, Hospital Management

*Corresponding Author : Je-Ran Chun(hitjr5000@naver.com)

Received March 14, 2021

Accepted May 20, 2021

Revised May 3, 2021

Published May 28, 2021

1. 서론

한국은 2020년 합계출산율이 사상 최저인 0.84명을 기록했다. 한국은 경제협력개발기구(OECD) 회원국의 합계출산율 비교에서도 2013년 이후 2018년까지 6년 연속 꼴찌를 기록했다. 한국은 OECD 회원국 중 유일하게 ‘출산율 0명대 국가’가 됐다[1].

이런 추세면 오는 2040년으로 예상됐던 인구 감소시기도 앞당겨질 것으로 보인다. 통계청은 2019년 발표한 장기인구추계에서 인구 감소가 시작되는 시기를 2032년에서 2029년으로 앞당겼다. 합계출산율이 1명 이하에 장기간 머물러 있을 경우 인구 감소가 더욱 가파르게 나타날 수 있다는 우려가 나온다. 결국, 정부와 지자체의 현금 지원만으로는 부부들이 아이를 더 낳도록 유도하기 어려운 상황이다. 따라서, 보육과 교육 정책을 대대적으로 전환해야 하고, 특히 외국인 인력 유입과 이들이 출산해서 자녀를 기를 수 있는 법적 제도적 환경을 구비하는 발상의 전환이 필요하다[2][3].

태아의 보건운영관리는 국가의 정책뿐만 아니라 병원의 보건관리시스템의 운영과 관리에 따라 출산에 상당하게 영향을 줄 수 있다. 이러한 보건운영관리는 저출산과 태아의 건강한 출산 그리고 산모의 건강에도 유의미한 정책으로 볼 수 있다. 즉, 병원에서 태아 보건을 위해 적절한 안정성과 신뢰할 수 있는 임신과 출산 그리고 육아라는 종합적인 구동축에서 종합적인 건강정보를 제공하거나 지원할 수 있는 병원관리 시스템이 구축되어야 한다.

UN의 SDGs(유엔 지속가능발전목표)는 빈곤, 질병, 여성, 아동, 난민 등의 인류의 보편적 문제와 기후변화, 에너지, 환경오염, 물 등의 지구 환경문제, 기술, 주거 노사, 고용, 생산 소비, 법, 대내외경제 등의 경제 사회문제를 2030년까지 17가지 주목표와 169개 세부목표를 해결하고자 하는 국제적인 공동목표로 삼고 있다. 이를 위해 지속 가능한 발전을 위해 전 연령대의 건강한 삶을 보장하고 복지를 증진시키는 것은 필수적이다[4][5].

SDGs의 세부목표는 2030년까지 예방 및 치료를 통한 비전염성 질병으로 인한 조기 사망을 3분의 1 감축시키고 정신 건강 및 복리 증진과 2030년까지 신생아 및 5세 미만 아동의 예방 가능한 사망을 없애고, 모든 국가는 신생아 사망을 1,000건의 생존 출산당 적어도 12건, 5세 미만 사망을 1,000건의 생존 출산당 적어도 25건으로 감축시키는 것을 포함한다. 2017년 출산 및 출산 중 사망자 수는 295,000명인 산모 사망률과 유사하다. 이러한 사망의 94%는 환경에 의해 발생하므로 대부분 예

방할 수 있다.

따라서 본 논문에서는 랜덤 포레스트(Random Forest)를 이용하여 Cardiotocograms(CTG) 검사에서 추출한 태아의 심박수, 태아의 움직임, 자궁 수축 등의 22개의 데이터 정보로 태아의 건강을 예측하는 모델을 제안하여 태아의 건강을 예측하고 사망을 예방할 수 있는 정확한 모델을 구축하는 것을 목표로 한다.

2. 관련 이론

2.1 배깅(Bagging)

배깅(Bagging)은 Bootstrap Aggregating의 약자로, 보팅(Voting)과 달리 동일한 알고리즘으로 여러 분류기를 만들어 보팅으로 최종 결정하는 알고리즘이다[6][7]. Fig. 1. 에서는 배깅에 대한 설명을 예시하고 있으며, 다음과 같은 방식으로 진행된다. (1) 동일한 알고리즘을 사용하는 일정 수의 분류기 생성 (2) 각각의 분류기는 부트스트래핑(Bootstrapping)방식으로 생성된 샘플데이터를 학습 (3) 최종적으로 모든 분류기가 보팅을 통해 예측을 결정한다.

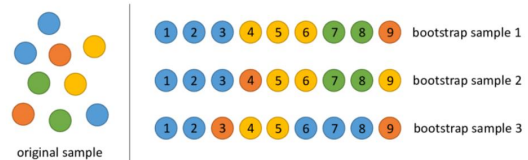


Fig. 1. Bagging

2.2 랜덤 포레스트(Random Forest)

기계 학습에서의 랜덤 포레스트(Random Forest)는 분류, 회귀 분석 등에 사용되는 앙상블 학습 방법의 일종으로, 훈련 과정에서 구성된 다수의 결정 트리로부터 분류 또는 평균 예측치(회귀 분석)를 출력함으로써 동작한다. 즉, 랜덤 포레스트란 여러 개의 의사결정트리를 만들고, 투표를 시켜 다수결로 결과를 결정하는 방법이다. 랜덤 포레스트는 여러 개의 결정 트리를 활용한 배깅 방식의 대표적인 알고리즘이다[8, 9].

배깅을 통한 랜덤 포레스트를 훈련시키는 과정은 다음과 같다. (1) 부트스트랩 방법을 통해 T개의 훈련 Data Set을 생성한다. (2) T개의 기초 분류기(트리)들을 훈련시킨다. (3) 기초분류기(트리)들을 하나의 분류기(랜덤 포

레스트)로 결합한다. 트리는 작은 편향(bias)과 큰 분산(variance)을 갖기 때문에, 매우 깊이 성장한 트리는 훈련데이터에 대해 과적합(overfitting)하게 된다. 부트스트랩 과정은 트리들의 편향은 그대로 유지하면서, 분산은 감소시키기 때문에 포레스트의 성능을 향상시킨다. 즉, 한 개의 결정 트리의 경우 훈련 데이터에 있는 노이즈에 대해서 매우 민감하지만, 트리들이 서로 상관화(correlated)되어 있지 않다면 여러 트리들의 평균은 노이즈에 대해 강인해진다. 포레스트를 구성하는 모든 트리들을 동일한 데이터셋으로만 훈련시키게 되면, 트리들의 상관성(correlation)은 커질 것이다. 따라서 배깅은 서로 다른 데이터셋들에 대해 훈련 시킴으로써, 트리들을 비상관화시켜주는 과정이다.

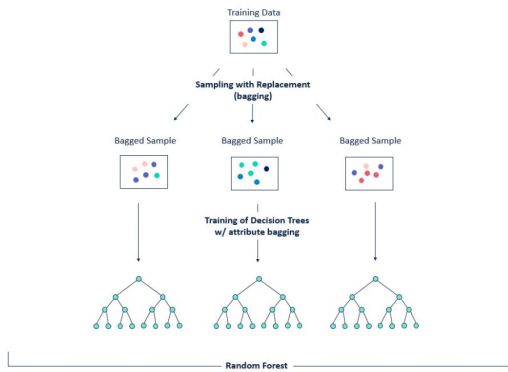


Fig. 2. Random Forest

2.2.1 주요 매개변수

랜덤 포레스트(Random Forest)의 가장 큰 영향을 주는 매개변수는 포레스트의 크기와 최대 허용 깊이 등이 있다. 포레스트의 크기는 총 포레스트를 몇 개의 트리로 구성할지를 결정하는 매개변수이다. 포레스트가 작으면 트리들을 구성하고 테스트하는데 짧은 시간이 걸리는 대신, 일반화 능력이 떨어져 임의의 입력 데이터 포인트에 대해 틀린 결과를 내놓을 확률이 높다. 반면에 포레스트의 크기가 크다면 훈련과 테스트 시간은 증가하지만, 포레스트의 결과값은 각 트리의 결과들에 평균을 취한 것으로 큰 포레스트의 결과값은 작은 포레스트보다 비교적 연속적이며 일반화 능력이 우수하다[10]. 최대 허용 깊이는 하나의 트리에서 루트 노드부터 종단 노드까지 최대 몇개의 노드를 거칠 것인지를 결정하는 매개변수이다. 최대 허용 깊이가 작으면 과소적합(underfitting)이 일어나고, 최대 허용 깊이가 크면 과대적합(overfitting)이 일

어나기 때문에 적절한 값을 설정하는 것이 중요하다[11].

2.3 분류결과표(Confusion Matrix)

분류결과표는 타겟의 원래 클래스와 모형이 예측한 클래스가 일치하는지 개수로 센 결과를 표로 나타낸 것이다. 정답 클래스는 행(row)으로 예측한 클래스는 열(column)로 나타낸다. 다음 Table.에서 대각선은 전체 중 해당 모델이 바르게 분류한 비율이다.

| | | Prediction | |
|--------|----------|------------|----------|
| | | Positive | Negative |
| Actual | Positive | TP | FN |
| | Negative | FP | TN |

Fig. 3. Confusion Matrix

2.4 분류 모델의 평가 지표(Classification report)

분류를 수행할 수 있는 기계 학습 알고리즘은 그 분류기의 예측력을 검증/평가하는 과정이 필요하다. 모델의 성능을 평가하려면 모델을 생성하기 전부터 애초에 데이터를 학습 세트와 평가 세트로 분리해서, 학습 세트로 모델을 만들고 평가 세트로 그 모델의 정확도를 확인하는 절차를 거친다.

Classification report는 분류 모델의 accuracy, recall, precision, f1-score 등을 포함한 머신러닝 분류 모델의 성능 평가 지표이다.

정확도(accuracy)는 전체 샘플 중 맞게 예측한 샘플 수의 비율을 뜻한다. 일반적으로 학습에 최적화 목적 함수로 사용된다.

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

정밀도(precision)은 양성 클래스에 속한다고 출력한 샘플 중 실제로 양성 클래스에 속하는 샘플 수의 비율을 말한다. FDS의 경우, 사기 거래라고 판단한 거래 중 실제 사기 거래의 비율이 된다.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

재현율(recall)은 실제 양성 클래스에 속한 표본 중에 양성 클래스에 속한다고 출력한 표본의 수의 비율을 뜻한다. FDS의 경우, 사기 거래라고 판단한 거래 중 실제 사기 거래의 비율이 된다.

$$recall = \frac{TP}{TP+FN} \quad (3)$$

F-점수(F-score)는 정밀도와 재현율의 가중조화평균(weight harmonic average)이다. 정밀도에 주어지는 가중치는 베타(beta)라고 한다. 베타가 1인 경우를 F1점수라고 한다.

$$F_{\beta} = \frac{(1+\beta^2)(precision \times recall)}{(\beta^2 \times precision + recall)} \quad (4)$$

3. 연구 방법 및 결과

3.1 학습 데이터

Table 1. 과 같이 Cardiotocograms(CTG) 검사에서 추출한 심박수, 태아의 움직임, 자궁 수축, 비정상 단기 변동, 단기 변동의 평균값 등의 22개의 정보를 데이터 세트로 구성된다. 각 정보들은 2126개의 데이터로 이루어져 있으며 정상, 의심, 병에 걸린 태아(1, 2, 3)로 분류된다. 본 연구에서는 이 데이터를 활용해 세 가지 태아의 건강 상태를 분류하는 예측 모델을 제안한다.

Table 1. Data set

| |
|---|
| 0: baseline value |
| 1: accelerations |
| 2: fetal_movement |
| 3: uterine_contractions |
| 4: light_decelerations |
| 5: severe_decelerations |
| 6: prolonged_decelerations |
| 7: abnormal_short_term_variability |
| 8: mean_value_of_short_term_variability |
| 9: percentage_of_time_with_abnormal_long_term_variability |
| 10: mean_value_of_long_term_variability |
| 11: histogram_width |
| 12: histogram_min |

| |
|--------------------------------|
| 13: histogram_max |
| 14: histogram_number_of_peaks |
| 15: histogram_number_of_zeroes |
| 16: histogram_mode |
| 17: histogram_mean |
| 18: histogram_median |
| 19: histogram_variance |
| 20: histogram_tendency |
| 21: fetal_health(class) |

3.2 데이터 분석

본 예측모델은 건강한 태아의 출산과 육아에 관한 국가의 보건정책과 병원의 보건운영관리에 적절한 기준과 건강한 출산에 긍정적인 영향을 줄 수 있도록 접근하고자 한다. 출산은 특정 개인의 사회적 지위와 경제적 능력에 따라 건강에 영향을 미치는 부분이 존재해서는 안된다, 경제적 능력이 부족한 저소득층이 의료 지출 여력이 부족해 출산과정에서 병원의 안정적인 운영정책 내에 포함되지 못한다면 결과적으로 열악한 의료서비스를 받게 되며, 이에 따라 지속적인 건강악화 현상이 반복될 수 있다.

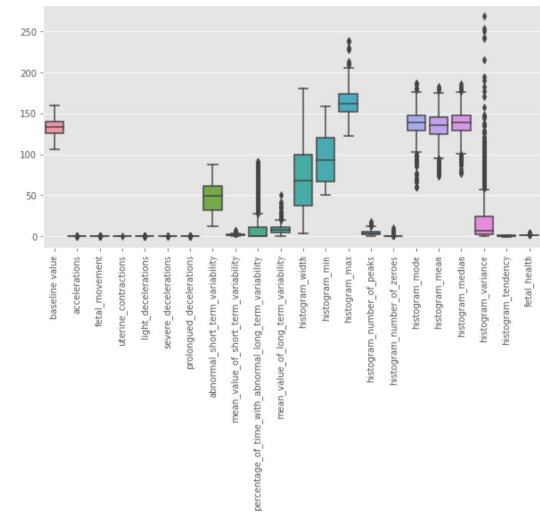


Fig. 4. Visualize data distribu

모델을 구축하기 위해서 전처리 과정에서 데이터를 분석한다. 본 논문에서는 21개의 항목과 1개의 클래스 정보를 처리한다. 각 정보와 클래스는 모두 2,126개의 데이터를 가진다. 그림 Fig. 4. 와 같이 데이터 분포를 시각화하여 분포가 불균형한 이상 데이터를 찾을 수 있다. 이 데이터들을 표준편차의 상한 및 하한의 임계값을 설정하여 이상 값을 제거하면 불규칙한 데이터로 인해 낮아질

수 있는 정확도를 높일 수 있다.

이상적인 분류를 위해서는 클래스들의 분포가 균일하게 이루어져야 한다. 하지만 데이터셋 대부분은 클래스의 분포가 균일하지 못하고 한쪽으로 치우치는 경향이 있다. 학습 알고리즘들은 클래스의 분포가 균일하다고 가정하기 때문에 클래스의 비율이 비슷하지 않은 경우 전체적인 데이터에 대해 학습하지 못하고 편향된 학습을 하게 된다. 그림 Fig 5. 와 같이 본 논문에서 사용하는 데이터 또한 클래스의 비율이 고르지 못하다. 따라서 리샘플링을 이용하여 클래스의 비율을 맞추었다.

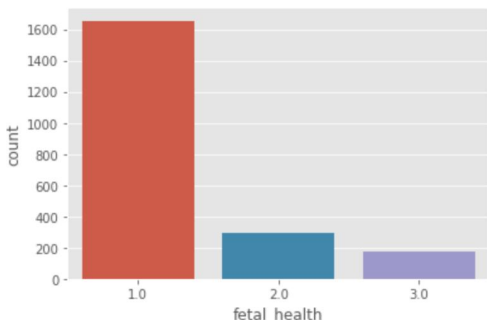


Fig. 5. Distribution of Target class

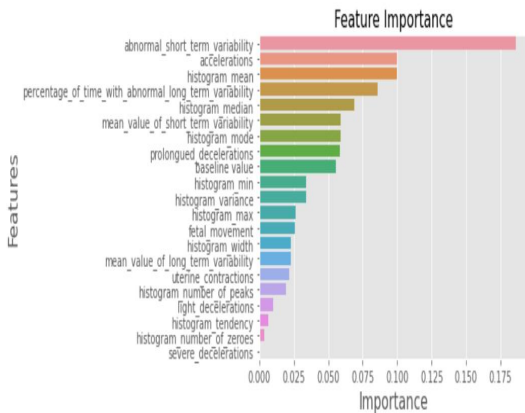


Fig. 6. Feture Importance

그림 Fig 6. 은 데이터 중요도에 따라 데이터 프레임 을 내림차순으로 정렬한 결과이다. 그래프를 보면 abnormal_short_term_variability(비정상 단기 변동)에서 가장 높은 중요도를 나타내고 있다.

3.3 결과

그림 Fig. 7. 의 분류결과표를 예서는 이 모델이 3개

클래스 모두 97%이상의 정확도를 보여준다.

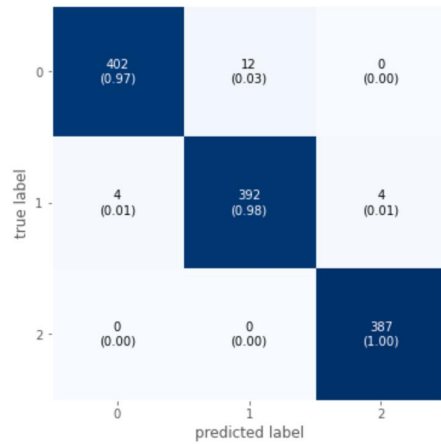


Fig. 7. Confusion matrix

Table 2. 와 같이 분류모델의 평가지표를 살펴보면 세 개의 클래스 모두 정밀도, 재현율, F-점수 등 95% 이상의 값을 보이며 정확도는 97.75%로 높은 수치를 나타내며 예측이 정확히 되고 있음을 확인할 수 있다.

Table 2. Classification Report

| | precision | recall | f1-score | support |
|------------|-----------|--------|----------|---------|
| 1.0 | 0.99 | 0.95 | 0.97 | 414 |
| 2.0 | 0.95 | 0.99 | 0.97 | 400 |
| 3.0 | 1.00 | 1.00 | 1.00 | 387 |
| accuracy | | | 0.98 | 1201 |
| macro avg | 0.98 | 0.98 | 0.98 | 1201 |
| weight avg | 0.98 | 0.98 | 0.98 | 1201 |

4. 결론

본 논문에서는 검출, 분류, 회귀 등 다양한 문제에 활용되는 랜덤 포레스트를 이용하여 태아의 건강상태를 분류하는 모델을 연구하였다. 태아의 건강상태를 나눈 클래스들의 비율이 불균형함에 따라 데이터 리샘플링을 이용하여 소수의 클래스를 복제하여 주된 클래스의 개수만큼 충족시킬 수 있었다. 데이터의 균형을 맞춘 후 정확도가 97.75%로 4-5% 향상된 결과를 얻을 수 있었다.

본 모델은 병원의 보건경영에 긍정적인 영향요인이 될 수 있으며, 보다 효과적인 태아 보건관리가 가능할 것으로 기대하고 있다. 특히 산전이나 출산 및 산후서비스를

제공할 수 있는 기본정보인프라로 볼 수 있으며 이러한 모델이 적극적으로 도입 운영된다면 태아보전에 대한 가이드라인을 제공할 수 있어서, 병원에서 운영하는 태아 관련 보건운영서비스에 경제적 그리고 효율적인 접근이 가능할 것이다. 이를 통해서 태아 보전에서 무엇보다 중요한 조산이나 유산 그리고 극히 일부이지만 태아의 사망과 산모의 사망 등도 최소화가 가능할 것이다. 따라서 병원의 효율적인 태아 보건운영관리를 통해서 이러한 위험에 대비하고 건강한 태아의 성장에 유용한 가이드라인을 제공할 수 있을 것이다.

이에 본 예측 모델을 통해 발생 할 수 있는 태아의 사망과 병을 사전에 정확히 예측하여 우선적으로 관리함으로써 태아의 사망과 병 예방에 기여할 수 있을 것이라고 기대한다.

REFERENCES

- [1] J. H. Oh. (2021). A study on prediction for reflecting variation of fertility rate by province under ultra-low fertility in Korea. *The Korean Journal of applied Statistics*, 34(1), 75-98.
- [2] D. C. Kim. (2018). Effect of the Child Benefit on the Increase of the Total Fertility Rate. *The Journal of Korean Public Policy*, 20(2), 3-25. DOI : 10.37103/kapp.20.2.1
- [3] S. M. Lee & H. R. Song. (2018). Proposal on the Definition of the New Social Indicator, the Effective Total Dependency Ratio, and its Application - Focusing on the Effects on Raising Fertility Rate, Migration and Extending Retirement Age. *Korea Journal of Population Studies*, 41(2), 91-116. DOI : 10.31693/kjps.2018.06.41.2.91
- [4] Alison Smale. (2018). What the SDGs mean. *UN Chronicle*, 55(2), 6-7. DOI : 10.18356/ca912345-en
- [5] Priya Priyadarshini & P.C. Abhilash. (2018). Sustainability science and research for attaining UN-SDGs. *Journal of Cleaner Production*, 184, 609-610. DOI : 10.1016/j.jclepro.2018.02.270
- [6] Hossein Foroozand & Steven Weijs. (2017). Entropy Ensemble Filter: A Modified Bootstrap Aggregating (Bagging) Procedure to Improve Efficiency in Ensemble Model Simulation. *Entropy*, 19(10), 520. DOI : 10.3390/e19100520
- [7] S Pavan Kumar Reddy & U Sesadri. (2013). A Bootstrap Aggregating Technique on Link-Based Cluster Ensemble Approach for Categorical Data Clustering. *INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY*, 10(8), 1913-1921. DOI : 10.24297/ijct.v10i8.1468
- [8] Carlos J. Mantas, Javier G. Castellano, Serafin Moral-Garcia & Joaquín Abellán. (2019). A comparison of random forest based algorithms: random credal random forest versus oblique random forest. *Soft Computing*, 23(21), 10739-10754. DOI : 10.1007/s00500-018-3628-5
- [9] Vikas Jain & Ashish Phophalia. (2020). M-ary Random Forest - A new multidimensional partitioning approach to Random Forest. *Multimedia Tools and Applications*. DOI : 10.1007/s11042-020-10047-9
- [10] Robert P. Sheridan. (2013). Using Random Forest To Model the Domain Applicability of Another Random Forest Model. *Journal of Chemical Information and Modeling*, 53(11), 2837-2850. DOI : 10.1021/ci400482e
- [11] Daniel S. Chapman, Aletta Bonn, William E. Kunin & Stephen J. Cornell. (2009). Random Forest characterization of upland vegetation and management burning from aerial imagery. *Journal of Biogeography*, 37(1), 37-46. DOI : 10.1111/j.1365-2699.2009.02186.x

전 제 란(Chun, Je Ran)

[정회원]



- 2005년 8월 : 청주대학교 경영학석사
- 2008년 8월 : 청주대학교 경영학박사
- 1997년 ~ 2008년 2월 : (의)효성병원 경영관리원장
- 2010년 3월 ~ 2020년 8월 : 대전보건대학교 의무행정정보학과 교수
- 2020년 9월 ~ 현재 : (의)효성병원 경영관리원장

- 관심분야 : 원부관리, 병원정보시스템, 병원고객관계관리, U-Health system, 보건운영관리
- E-Mail : hitjr5000@naver.com