ISSN: 3022-5388
JKAIA website: https://accesson.kr/jkaia
DOI: http://dx.doi.org/10.24225/jkaia.2023.1.1.11

머신러닝 알고리즘 기반의 의료비 예측 모델 개발

Development of Medical Cost Prediction Model Based on the Machine Learning Algorithm

Han Bi KIM[1], Dong Hoon HAN[2]

Received: April 24, 2023. Revised: May 30, 2023. Accepted: June 30, 2023.

Abstract

Accurate hospital case modeling and prediction are crucial for efficient healthcare. In this study, we demonstrate the implementation of regression analysis methods in machine learning systems utilizing mathematical statics and machine learning techniques. The developed machine learning model includes Bayesian linear, artificial neural network, decision tree, decision forest, and linear regression analysis models. Through the application of these algorithms, corresponding regression models were constructed and analyzed. The results suggest the potential of leveraging machine learning systems for medical research. The experiment aimed to create an Azure Machine Learning Studio tool for the speedy evaluation of multiple regression models. The tool faciliates the comparision of 5 types of regression models in a unified experiment and presents assessment results with performance metrics. Evaluation of regression machine learning models highlighted the advantages of boosted decision tree regression, and decision forest regression in hospital case prediction. These findings could lay the groundwork for the deliberate development of new directions in medical data processing and decision making. Furthermore, potential avenues for future research may include exploring methods such as clustering, classification, and anomaly detection in healthcare systems.

Keywords : Machine learning, Algorithm, Models, Regression, Prediction, Azure Mchine Learning Studio, R, Error, Accuracy, Performance metrics, Health care

Major Classification Code : Artificial Intelligence, Machine Learning, Prediction Analysis

1. Introduction

오늘날 세계에서는 주로 코로나바이러스 감염증-19 로 인한 전염병 문제, 양질 의료에 대한 수요, 인구 노령화, 만성질환, 장기 질환 등을 앓고 있는 사람들의 증가로 의료비가 급격히 상승하는 추세이다. 과도한 의료비 지출을 막기 위해서는 본인의 의료비를 직접 예측하여 과납을 막고 그에 상응하는 보험금을 납부할 수 있어야 하며, 빠른 의료비 지출 증가세 속에서 높은 의료비를 지불하는 상황에 대비할 필요가 있다. 또한 환자가 자신의 건강에 대한 보다 종합적인 관점을 갖고, 생명과 관련된 상황에 대한 자신의 통계 자료를 모니터링하기 위해 자신의 의료

1 First Author. Undergraduate, Big Data Medical Convergence, Eulji University, South Korea. Email: khb1022@g.eulji.ac.kr
2 Corresponding Author or Second Author. Researcher, Medical Artificial Information Center, South Korea. Email: d555v@naver.com

ⓒ Copyright: The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://Creativecommons.org/licenses/by-nc/4.0/) which permits unrestricted noncommercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

데이터에 접근할 수 있는 것이 중요하다. 의료 데이터 자원에 대한 접근은 의료 서비스 제공자, 투자자, 정부 기관, 연구원, 과학자 등의 효과적인 활동을 위해서도 필요하다. 오늘날 컴퓨터 과학 기술 발전은 이러한 필요성을 충족시킨다. 그러나 이 방법으로는 데이터 전송 및 처리에 대한 세계적 표준을 개발하는 데 문제가 있다. 의료 기술의 급속한 발전과 관련한 이유로, 이러한 표준의 확립이 점점 더 시급해지는 추세이다. 이 프로젝트의 가장 중요한 기본 원리 중 하나는 헬스케어 분야의 데이터를 쉽고 편리하게 수집, 교환, 분석할 수 있는 통합된 정보 및 디지털 헬스케어 시스템을 개발하는 것이다(World Economic Forum, 2018).

최근 몇 년 동안 통계적 방법을 적용한 의료 데이터 처리는 점점 인기를 얻고 있으며 많은 전문가들의 연구 주제였다. 이는 정보와 디지털 기술의 급속한 발전과 엄청난 양의 헬스케어 분야 데이터 생성의 결과이다. 빅데이터 처리의 필요성, 처리를 위한 방법 선택, 그리고 현대 의학에서 이러한 목적을 위한 머신러닝과 데이터의 사용은 이 분야의 연구자들이 혁신적인 탐구를 할 수 있도록 기여한다.

헬스케어 분야의 목표가 질병의 위험을 예측하고 적절한 예방을 보장하며 인간의 특정 요구를 충족시키기 위해 치료 계획을 정확하게 하는 것이라면, 상당수의 환자로부터 얻은 포괄적인 데이터에 접근하는 것이 전제조건이다. 또한 의료 데이터의 표준을 정하는 것이 인간 중심의 헬스케어 시스템을 만드는 핵심이다. 표준화 과정은 기술적인 면에서 도전적일 수도 있지만, 보건 협력의 목표 달성을 위해서는 매우 중요한 방법이다. 뿐만 아니라 의료 분야의 정보 및 디지털 기술의 발전은 의료 지식의 발전에 기여한다는 점에서 메우 중요한 역할을 한다.

2. Literature Review and Discussion

2021 년 미국에서는 15 억 원이라는 코로나 치료비 폭탄을 맞은 50 대 남성의 사례가 언론에 보도된 바 있다(Jung, 2021). 또한 보험연구원에 따르면, 2019 년을 기준으로 지난 5 년간 국내총생산(GDP) 대비 의료비 지출에 대한 증가세가 OECD 국가 중 미국이 0.6%, 일본은 0.3%였던 반면 우리나라가 1.5%로 가장 빨랐다는 연구 결과도 존재한다(Kim, 2021).

Boston Consulting Group(BCG)과 협력하여 이행된 세계경제포럼 프로젝트 '헬스케어의 가치'는 지난 3 년간 의료 시스템을 개선하는 사고와 연구의 중요한 원천이 되었다. 2016 년 7 월 프로젝트 착수 이후, 프로젝트 이해관계자와의 협력을 기반으로 헬스케어 시스템 변화를 위한 주요 대비책이 강구되었다. 2017 년 헬스케어 시스템 개선의 핵심사항에 대한 보고서는 환자에 대한 진정한 가치보다는 주로 기회와 프로세스에 대한 지침에 초점을 맞추고 있는 헬스케어에 대한 접근방식을 언급하며, 환자에게 중요한 시스템의 기준은 의료 서비스의 양이 아닌 품질과 비용이고, 헬스케어 시스템 변화의 기반이 되는 구성 요소는 컴퓨터 과학, 지불 비용, 연구 및 도구 지표, 의료 서비스 계획이라고 기술한다(World Economic Forum, 2017). 이 프로젝트에 기술된 가치에 기반을 둔 접근방식은 특정 인구 집단과 관련된 결과를 측정하는 것부터 시작하였다. 측정 결과는 그 뒤에 치료량을 늘린 것이 아닌, 의료 서비스의 품질 향상과 비용 절약에 근거한 환자 치료 계획을 통해 해당 인구 집단에 대한 의료 서비스를 조정하는 데 사용된다. 이러한 경우에 사람의 요구에 맞게 의료 서비스 제공 방법을 조정하고 공급자, 약사, 의료 파트너의 서비스 체계에서 협력하는 데 중점을 두었다.

3. Body

건강보험회사는 수혜자의 의료비에 지출하는 것보다 더 많은 돈을 모아야만 돈을 벌 수 있다. 한편, 일부 인구 집단에 대해 더 일반적인 조건이 존재하더라도 대부분의 돈은 환자의 희귀한 조건에서 발생하기 때문에 의료비는 예측하기 어렵다. 이 연구의 목적은 나이, 체질량지수, 흡연 여부 등 사람들의 데이터를 기반으로 보험비를 정확하게 예측하는 것이다. 추가로, 우리는 보험비에 영향을 미치는 가장 중요한 변수가 무엇인지 또한 결정할 것이다. 이 추정치는 예상되는 치료비에 따라 연간 보험료의 가격을 더 높게 또는 더 낮게 설정하는 보험료표를 만드는데 사용될 수 있다.

본 연구에서는 Kaggle 플랫폼에서 공개적으로 사용할 수 있는 2018 년 2 월에 얻은 "Medical cost personal datasets"의 데이터를 사용했다. 표 1 의 구조 하에 만들어진 데이터 세트는 CSV 형식으로 7 열, 1338 행으로 구성되어 있다. 데이터의 다중 공선성을 테스트했을 때, 없는 것으로 확인되었고, 데이터 세트에 대한 내용은 표 2 에 설명되어 있다.

이것은 회귀 문제이며, 가장 널리 사용되는 통계 방법의 하나다. 회귀 알고리즘 기반의 머신러닝 모델을 구축하기 위해 Microsoft Azure Machine Learning Studio 를 사용했다. Microsoft Azure ML Studio 는 데이터 세트와 분석 모듈을 포함하는 시각적 개발 환경이다. 이 환경은 예측 데이터 분석 솔루션을 생성, 테스트 및 배포하도록 설계된 공유 기능을 지원하고, 다음과 같은 문제점을 해결할 수 있다(Kang, 2021):

1) 학습을 위한 컴퓨팅 자원의 유연성 부족
2) 머신러닝을 위한 GPU 기반 환경 구성의 어려움
3) 학습에 필요한 도구 설치 및 설정의 어려움
4) 실험 기록 및 버전화의 어려움

Table 1: Dataset structure

age	sex	bmi	children	smoker	region	charges
19	female	27.90	0	yes	southwest	16884.92
18	male	33.77	1	no	southwest	1725.552
23	male	33.00	3	no	southwest	4449.462
33	male	22.71	0	no	northwest	21984.47
32	male	28.88	0	no	northwest	3866.855
31	female	25.74	0	no	southwest	3756.622
46	female	33.44	1	no	southwest	8240.59
37	female	27.74	3	no	northwest	7281.506
37	male	29.83	2	no	north west	6406.411
...	,,,	,,,	,,,	,,,	,,,	,,,

Table 2: Dataset description

Data field	Content	Data type
age	age of the main beneficiary	numeric
sex	gender of the insurance contractor	string
bmi	body mass index	numeric
child	number of children on health insurance number of dependents	numeric
smoker	smoking	string
region	US resident residential area	string
charges	individual medical expenses accruedunder health insurance	numeric

4. Model Construction and Experiment

우리는 회귀변수 Y 로 charges(의료 비용)를 선택하고, 설명변수로는 X1 - age, X2 - sex, X3 - bmi, X4 - children, X5 - smoker, X6 - region 을 선택하고 X 가 Y 에 미치는 영향에 관해 연구했다. charges 가 age 와 bmi 에서는 두 자리가 최대지만, 다른 변수들에서는 만 단위까지 올라가는 극명한 차이가 존재한다. 보통 단위 차이가 큰 변수로 인해 학습이 제대로 이루어지지 않을 경우를 대비하여 스케일링 작업을 수행하지만 앞서 말했듯이 charges 는 종속변수이기 때문에, 스케일링이 필요하지 않았다.

R 프로그래밍을 통해 데이터 분포를 살펴본 결과 나이가 20 대보다 어린 사람들의 수가 매우 많으며, 그로 인해 자녀의 수 역시 0 에 매우 많이 몰려있음을 알 수 있다. 따라서 이러한 분포에 맞춰 훈련과 테스트세트를 나누는 것이 중요해 보인다. 남녀 비율은 거의 비슷하며, bmi 의 경우 약간은 왼쪽으로 치우친 분포를 보인다. 흡연자와 비교하면 비흡연자의 비중이 높은데, 이 역시 나이와의 관련성을 의심해볼 수 있다.

복합적 box plot 을 통해 회귀변수와 설명변수 간의 관계를 살펴본다. age-charges 에서 나이의 평균이 증가함에 따라 의료비 또한 증가함을 확인할 수 있고, sex-charges 에서는 남성의 수치가 여성의 수치보다 다소 높을 뿐, 큰 차이를 보이지 않는다. smoker-charges 에서 비록 비흡연자의 경우에 흡연자보다 높은 경우가 발생하지만, 흡연자인 경우가 비흡연자보다 의료 비용이 더욱 발생한다는 사실을 알 수 있다. 더불어, 흡연자의 최소 의료비가 비흡연자의 평균보다 높다는 사실에 초점을 둘 수 있겠다.

따라서, 비용(charges) 분포를 다시 그려보고, 흡연 여부(smoker)로 분류해본다. 흡연 여부에 따라 나이, 체질량지수, 자녀의 수별 의료 비용을 분석해보겠다. 앞서 말한 세 변수와 의료 비용이 함께 증가함에도 불구하고, 흡연 여부가 의료 비용에 가장 높은 영향을 미치는 것으로 보인다.

변수 간 상관관계는 그림 1 에서 알 수 있듯이, smoker-charges 를 제외하고는 변수들 사이에 상관관계가 거의 없음을 확인할 수 있었다.

Figure 1: Correlation by variable

표 3 에 제시된 모든 Microsoft azure 알고리즘은 데이터 세트에 대한 학습을 통해 통제되는 머신러닝 알고리즘이다. 표 2 에 제시된 회귀 구성 알고리즘을 기반으로 한 머신러닝 모델의 구성은 그림 5 에 제시되어 있다.

그림 2 에서는 데이터 세트 블록으로 추가 가공을 위해 데이터를 로드한다. 메타데이터 편집 블록을 사용하면 데이터 세트에서 열을 선택하여 메타데이터를 편집할 수 있다. 데이터 분할 블록은 데이터 세트의 행을 훈련과 모델 테스트를 위한 두 가지 다른 집합으로 나눈다. 사용된 알고리즘은 베이지안 선형 회귀, 신경망 회귀, 향상된 의사결정 트리 회귀, 선형 회귀, 의사결정 포레스트 회귀 블록은 위의 알고리즘을 사용하여 회귀 모형을 생성한다. 회귀 모형의 훈련은 훈련 데이터 세트의 모델 훈련 블록을 사용하여 구현되었다. 개발된 회귀 모형은 테스트 부분의 모델 스코어링 블록을 사용하여 테스트하였다.

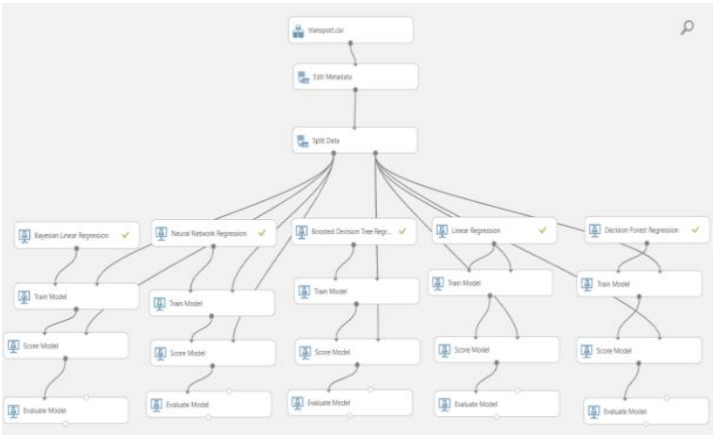

Figure 2: Implemented Azure ML Algorithm

Table 3: Azure ML Studio Regression Algorithm Used

Algorithm	Characteristic	Accuracy	Training time
Bayesian linear regression	based on clarifying the probabilities of hypotheses using Bayesian rules. Initial data are consistent with probability functions for the formation of parameter estimation..	-	moderate
Linear regression	Establishing a linear relationship between one or more independent variables and a numeric result or dependent variable.	-	quick
Neural network regression	Built on a statistical model using adaptive scaling and approximation of non-linear input functions.	high	-
Boosted decision tree regression	Used to generate regression tree ensembles via deployment. Learned by establishing the remainder of the tree that came before. Therefore, deployment in decision tree ensembles tends to increase accuracy, but may come with some small range risks.	high	moderate
Decision forest regression	It consists of a decision tree ensemble. Each tree in the decision forest produces a Gaussian distribution with predictions. In a tree ensemble, an aggregation is performed on the tree ensemble to find the Gaussian distribution that is closest to the combined distribution of each tree.	high	moderate

5. Results

구축된 모델은 평균 절대 오차(MAE), 평균 제곱 오차(RMSE), 상대 절대 오차(RAE), 상대 제곱 오차(RSE), 결정계수로 평가했다. 평가 지표의 의미는 다음과 같다:

1) 평균 절대 오차(Mean Absolute Error)는 예측이 실제 결과에 얼마나 가까운지를 측정한다. 점수가 낮을수록 모델의 예측이 더 정확하다.

2) 평균 제곱 오차(Root Mean Square Error)는 모형의 오차를 요약하는 단일 값을 생성한다. 차이를 제곱하면 과대 예측과 과소 예측의 차이를 무시한다. 제곱근을

취하여 원래 오차의 척도로 다시 돌리는데, 점수가 낮을수록 모델의 예측이 정확하다.

3) 상대 절대 오차(Relative Absolute Error)는 기댓값과 실제 값 사이의 상대적 절대 차이다. 평균 차이를 산술 평균으로 나누기 때문에 상대적인 척도를 제공하며, 값이 작을수록 모델의 예측이 더 정확하다.

4) 상대 제곱 오차(Relative Squared Error)는 실제 값의 총 제곱 오차를 나누어 예측값의 총 제곱 오차를 유사하게 정규화한다. 작은 값이 더 나은 모델이며, 예측이 실제 값과 얼마나 가까운지에 대한 상대적인 척도를 제공한다.

5) 결정계수(Coefficient of Determination): 모델이 종속 변수의 변동을 얼마나 잘 설명하는지를 측정하는 지표이다. 값이 1 에 가까울수록 모델이 데이터를 잘 설명하고 있으며, 0 에 가까울수록 모델의 설명력이 낮다.

표 4 에 의하면, 평균 절대 오차에 따라 의사결정 포레스트 회귀 분석과 향상된 의사결정트리 회귀 분석을 사용했을 대 실제 결과 데이터 예측과 가장 근접하게 구축되었고, 신경망 회귀 분석을 사용했을 때는 최악의 결과를 얻었다. 선형 회귀 분석과 베이지안 선형 회귀 분석의 구축 결과에서는 거의 동일한 수준에서 더 높은 평균 절대오차를 얻었다.

획득된 평균 제곱 오차값 또한 의사결정 포레스트 회귀및 향상된 의사결정트리 회귀의 우수성을 나타낸다. 선형 회귀 및 베이지안 선형 회귀 값은 앞에서 추정한 의사결정 포레스트 회귀 및 향상된 의사결정 트리 회귀의 평균 제곱 오차 값보다 크며 거의 동일한 수준이다. 평균 제곱 오차 추정치 중 가장 최악의 값은 신경망 회귀 알고리즘이었다.

상대 절대 오차와 상대 오차 제곱 측면에서 회귀 모델링 알고리즘의 추정 우선순위는 앞선 두 결과와 동일하다.

결정계수의 가장 높은 값은 향상된 의사결정트리 회귀 알고리즘을 기반으로 구축된 모형을 테스트하여 얻은 것으로, 의사결정 포레스트 회귀로 얻은 결정계수의 값은 그보다 약간 낮다. 결정계수는 독립변수의 영향으로 인한 종속 변수 분산의 일부, 즉 변화의 크기를 결정한다. 이 결정계수 값에 기초한 회귀 모형은 적합한 것이라고 할 수 있다. 선형 회귀 분석과 베이지안 선형 회귀 분석을 사용하여 구한 모형도 적합한데, 이는 결정계수가 각각 76.9%와 74.4%이기 때문이다.

따라서 위의 알고리즘에 따라 개발된 회귀 모형을 테스트한 결과, 실제 데이터에 가장 근접한 모델은 향상된 의사결정트리 회귀 알고리즘에 기반한 회귀 모형이며, 의사결정 포레스트 회귀 알고리즘을 기반으로 구축된 머신러닝 모델 또한 근접한 것으로 추정된다. 선형 회귀와 베이지안 선형 회귀 모형은 보다 낮은 값을 얻을 수 있었다.

신경망 회귀 알고리즘을 사용하여 구축된 모델은 부적절하므로 사용할 수 없다. 이 결과는 이 알고리즘을 적용하기 위해서는 추가적인 설정, 신중한 데이터 세트의 사전 준비가 필요하다는 것을 의미한다. 신경망에 기반한 다른 회귀 알고리즘 사용을 추천할 수도 있다.

Table 4: Evaluation of developed regression models by error indices (rounded to the second decimal place)

Algorithm	MAE	RMSE	RAE	RSE	Coefficient of Determination
Bayesian linear regression	4200.31	6147.12	0.46	0.26	0.74
Neural network regression	16033.19	17227.07	1.75	2.01	-1.01
Boosted decision tree regression	504.15	713.78	0.06	0.00	1.00
Linear regression	4124.48	5769.02	0.46	0.23	0.77
Decision forest regression	1392.57	2419.91	0.16	0.04	0.96

6. Conclusion

수리통계학 및 머신러닝을 기반으로 데이터를 분석한 뒤, 모델을 적용하는 과정에서 적절한 회귀모델을 구축하고 이러한 알고리즘을 토대로 얻은 결과를 기반으로 분석을 수행하였다. 연구 결과는 의료 연구에서의 머신러닝 시스템 활용 가능성을 시사한다. 또한, 제시된 방법들은 이 영역에서 의료 데이터 처리 및 의사결정의 새로운 영역을 전략적으로 개발하는 데 기초가 될 수 있을 것이다.

데이터 분석의 통계적 방법은 의료 기술의 비용 효율성을 높이는 증거 기반 의학의 신뢰할 만한 도구이다.

사람에 초점을 맞춘 통합 정보 및 디지털 건강 시스템 구축의 주된 목표는 환자 치료의 근본적인 모델과 의료의 혁신적인 가치를 지향하는 모델을 지원하는 것이다. 의료 결과 및 비용에 대한 데이터의 지속적인 분석이 인구의 주요 부분을 보다 정확하게 식별하고 이러한 부분의 기능을 최적으로 보장하기 위한 개별 조치를 개발하는 의료 시스템의 체계적 개선을 위한 벡터를 식별해야 한다. 이러한 지속적인 개선 주기를 지원하기 위한 정보 인프라는 표준에 기반하여 다양한 데이터 수집, 통합, 표현, 접근이 가능한 데이터 시스템 및 아키텍처의 개발을 필요로 한다. 이러한 조치는 의료 시스템과 다른 이해관계자 간의 데이터 교환을 가능하게 하여 혁신을 가속화할 수 있도록 할 것이다. 또한 통합된 정보 및 디지털 의료 시스템 구축의 필요성은 병원, 학계, 임상의 및 연구원의 요구에 기인한다.

보험, 제약, 의료 및 분석 회사는 대규모의 구조화·반구조화·비구조화된 데이터 세트에 대한 지적인 분석을 수행한다. 건강 정보학 표준의 개발은 또한 의료 과학의 임상 방법론, 연구 및 개발을 위한 데이터 세트를 기반으로 지식을 생성하는 정교한 분석 도구(예: 자동화된 의사결정 지원 도구)의 개발을 가속화할 것이다.

본 논문에서 획득한 결과를 통해 의료 보험료 변화의 추세와 개개인의 의료비 특성을 파악하여 맞춤형 보험 상품을 기획할 수 있을 것이며, 의료 시스템에서 클러스터링, 분류, 이상 탐지 등을 위한 머신러닝 시스템의 사용에 대한 추가 연구를 기대한다.

References

Alpenberg. J. & Scarbrough. D. P. (2015). *Lean Healthcare and Ontario Case Costing* - An Examination of Strategic Change and Management Control Systems (2014). CAAA Annual Conference. http://dx.dio.org/10.2139/ssrn.2538388

Botchkarev. A. (2018). *Evaluating Hospital Case Cost Prediction Models UsingAzure Machine Learning Studio*. https://doi.org/10.48550/arXiv.1804.01825

Choi, M. R. (2018). *Medical cost personal datasets*. https://www.kaggle.com/datasets/mirichoi0218/insurance.

Jung. Y. S. 15억 원 코로나 치료비 폭탄 맞은 미 50대… 보험 있어도 불감당. 연합뉴스. (2021년 2월 9일), https://www.yna.co.kr/view/AKR20210209088800075.

Kang, M. S. (2021). *Machine Learning: Concepts, Tools and Data Visualization*. World Scientific, p. 80.

Kim. E. Y. 우리나라 의료비 지출 증가세, OECD 국가 중 '가장 빨라'. (2021년 4월 8일), https://www.docdocdoc.co.kr/news/articleView.html?idxno=2009477.

Klochko, O. V., Gurevych, R. S., Nagayev, V. M., Dudorova, L. Yu. & Zuziak, T. P. (2022). *Data mining of the healthcare system based on the machine learning model developed in the Microsoft azure machine learning studio*. J. Phys: Conf. Ser. 2288 012006

Microsoft. (2022). Microsoft Azure Machine Learning Studio. https://ml.azure.com/home.

R Studio. https://posit.co/download/rstudio-desktop.

Wang. C. W. (2008). *New ensemble machine learning method for classification and prediction on gene expression data.* Encyclopedia of healthcare information systems (IGI Global) pp 982–989.

World Economic Forum. (2017). *Value in healthcare: laying the foundation for health system transformation*. Cologny/Geneva, Switzerland.

World Economic Forum. (2018). *Value in healthcare: accelerating the pace of health system transformation*. Cologny/Geneva, Switzerland.

World Economic Forum. (2018). *Value in healthcare: mobilizing cooperation for health system transformation*. Cologny/Geneva, Switzerland.