applied sciences

MDPI

Article

A Machine Learning Approach to Predict Customer Usage of a Home Workout Platform

Qiuying Chen and SangJoon Lee *

Interdisciplinary Program of Digital Future Convergence Service, Chonnam National University, Gwangju 61186, Korea; chenqiuying11@gmail.com
* Correspondence: s-lee@jnu.ac.kr; Tel.: +82-62-530-1447

Abstract: Health authorities have recommended the use of digital tools for home workouts to stay active and healthy during the COVID-19 pandemic. In this paper, a machine learning approach is proposed to assess the activity of users on a home workout platform. Keep is a home workout application dedicated to providing one-stop exercise solutions such as fitness teaching, cycling, running, yoga, and fitness diet guidance. We used a data crawler to collect the total training set data of 7734 Keep users and compared four supervised learning algorithms: support vector machine, k-nearest neighbor, random forest, and logistic regression. The receiver operating curve analysis indicated that the overall discrimination verification power of random forest was better than that of the other three models. The random forest model was used to classify 850 test samples, and a correct rate of 88% was obtained. This approach can predict the continuous usage of users after installing the home workout application. We considered 18 variables on Keep that were expected to affect the determination of continuous participation. Keep certification is the most important variable that affected the results of this study. Keep certification refers to someone who has verified their identity information and can, therefore, obtain the Keep certification logo. The results show that the platform still needs to be improved in terms of real identity privacy information and other aspects.

Keywords: home workout; platform; machine learning; prediction; customer usage

Citation: Chen, Q.; Lee, S. A Machine Learning Approach to Predict Customer Usage of a Home Workout Platform. *Appl. Sci.* **2021**, *11*, 9927. https://doi.org/10.3390/app11219927

Academic Editor: Federico Divina

Received: 9 September 2021
Accepted: 21 October 2021
Published: 24 October 2021

1. Introduction

Digital technologies have profoundly and intensively changed social life, including transforming people's sports activities and health behaviors. Home workout clubs and influencers make it possible for followers and members to practice fitness activities in almost any place and time. In addition, online platforms demonstrate lifestyle sports and fitness tendencies and popularize messages associated with health and fitness [1]. Several accounts elaborately spread the experiences in applying fitness and physical activity applications or wearable devices. Others evaluate technical contributions to the level of such physical activities or investigate gender and age difference in the adoption of fitness applications [2].

To mitigate the transmission of the coronavirus disease (COVID-19) and reduce the interactions between unrecognized infected individuals and non-infected individuals, many measures have been applied, including quarantine, local confinement, lockdown, and isolation. Therefore, people have no access to places such as fitness studios, sports clubs, swimming pools, and other public facilities in this special period of time. In addition to the sedentary habits people may have, the reduction of outdoor daily activities and restriction to the indoor environment at home leads to a reduction in the level of physical activity. In this regard, the World Health Organization and health authorities have recommended digital tools for home-based sport and exercise routines to maintain a fundamental level of physical activity during the pandemic [3,4].

As indicated by a global study, professionals from the health and fitness industry consider "online fitness" as the leading fitness tendency in 2021 [5]. Influenced by COVID-

19, home workout has become popular as it enables individuals to easily practice activities at home without being limited by space and time. However, studies have indicated that the adoption of digital sports activities during the pandemic remains insufficient. Therefore, studies have been conducted to determine the factors that affect the practice of home workouts using big data analysis.

Machine learning is a sub-discipline of artificial intelligence (AI) in which computers derive new information or decisions by learning data using algorithms and programs, similar to human learning [6]. To obtain good discriminant results, an appropriate selection of variables is necessary, but the importance of certain variables remains unclear. In general, to exercise at home, an individual needs a timer or watch to time the activity or an application such as Keep. Keep is a sports application commonly used in China that has recently announced that it reached 100 million registered users, becoming the first Chinese sports application to reach this mark and China's largest social platform for sports [7]. This application integrates big data and AI with a focus on sports and employs science and technology to improve users' sports experiences and needs. Thus, this study aimed to find effective methods to determine whether users participated in an exercise in a given month, focusing on the home workout application Keep and using big data to propose practical solutions with a new approach.

Advanced analysis based on machine learning has become increasingly popular because it can provide business advantages for almost every industry. 'Shelter in place' and 'lockdown' orders implemented to minimize the spread of COVID-19 reduced opportunities to be physically active. For many, the home environment was the only viable option to practice physical exercises. Although COVID-19 has accelerated the development of the home workout industry, there are few studies evaluating the use of machine learning in this field. Therefore, we adopted a machine learning approach to predict whether home workouts will continue to be used based on the following reasons. First, compared to traditional business intelligence solutions, machine learning enables organizations to obtain more insights from structured and unstructured data. Second, compared with other analytical methods, machine learning provides multiple advantages for information technology (IT) personnel, data scientists, various business teams, and organizations. In the long run, the personnel cost of machine learning is lower than that of traditional analyses. Once machine learning is functioning normally, the predictive model can be adjusted by itself, which implies that only a small amount of manpower is required to complete accurate and reliable adjustments. Compared with traditional analysis methods, machine learning can solve problems and reveal insights faster and more easily, indicating operating plans for similar home workout platform companies and providing a competitive advantage.

The objectives of this study are summarized as follows. First, machine learning algorithms were applied to expand research methods in a continuous participation factor analysis. Second, we analyzed the factors influencing continuous exercise participation by Keep users. Third, we conducted a more scalable study of the analysis results to identify realistic meanings and present effective solutions.

The remainder of this paper is organized as follows. Section 2 introduces the use of machine learning to determine whether to participate in continuous exercise centered on Keep. To this end, we collected data in association with 7734 Keep users, as described in Section 3. Based on the collected learning data, in Section 4, we describe the implementation of a discriminant model using four supervised learning algorithms: support vector machine (SVM), k-nearest neighbor (KNN), random forest, and logistic regression. In Section 5, we evaluate the test sample discrimination and the variables that affect the determination of continuous participation. In Section 6, we identify the main variables of continuous user discrimination and extend the research discussion by analyzing the precision of discrimination.

2. Theoretical Background and Previous Studies

2.1. Home Workout in the Untact (Un-Contact) Era

The emergence of COVID-19, a highly contagious virus, has led to social distancing to prevent transmission. Social distancing is a type of infection control, which aims to decelerate the spread of the disease and ultimately minimize the mortality rate by reducing the likelihood of contact between infected and non-infected people [8]. In this regard, South Korea has reshaped its economy around a concept called "untact". It is a portmanteau created by adding the prefix "un"—used to denote negation in the English language—to the word "contact" [9]. Untact is a consumption tendency featuring anonymity and accessibility. As a result, untact consumption, untact marketing, and untact services are spreading, led by millennials who feel burdened by face-to-face services. As an example, home-based exercise programs are a feasible strategy to reduce the inactivity-induced losses in physical activity, as well as health- and skill-related components, in older adults [10].

COVID-19 is one of the biggest health challenges that the world has ever faced. Public health policy-makers need reliable predictions of confirmed cases in the future to plan medical facilities. In this regard, machine learning methods learn from historical data and make predictions about the events [11]. The bibliometric analysis in this study presents the most influential references related to COVID-19 from 1 December 2019 to 20 April 2020 and can be useful to improve the understanding and management of COVID-19 [12]. It summarizes and analyzes the evolution of the immediate impact of the COVID-19 pandemic on scientific production. In addition, it is important to recognize the interdisciplinary usefulness involving medical science, social science, engineering, and economics to solve problems [12]. In machine learning, models are built using historical data and are used to predict new outcomes. Regression, classification, clustering, and deep learning are some of the machine learning methods that have been successfully used in various domains such as image analysis, speech recognition, and health informatics. In [13,14] are some of the machine learning methods which have been successfully used in various domains such as image analysis, speech recognition, health informatics, etc. Many applications of machine learning methods for COVID-19 have been proposed, such as diagnosis and prognosis, patient outcome prediction, tracking and prediction of outbreaks, drug development, vaccine discovery, and false news prediction [15–19].

Home workout is defined as exercise done using one's weight, without specific tools, or muscle-building exercises using simple weight training equipment at home. With the development of devices and technologies such as computers and smartphones, the number of people practicing home workouts has increased; this is because sports information becomes available through videos with professional trainers without face-to-face contact [20]. Business wire reported that the global information technology training market was $64.6 billion in 2018, grew at an annual rate of 6% between 2019 and 2020, and is expected to reach $91.7 billion by 2024 [19]. In March 2020, 495,000 videos were searched on Google using the keyword "home workout". This figure is 663 times higher than the 781 times in 2010. In addition, when searching using the keyword "home workout", a total of 1.18 billion images and 388 million videos related to "yoga" were retrieved. As of May 2019, users could find 240 paid and free applications on sale using the keyword "home workout" on the Google App Store [21]. The data also demonstrate that home workout is growing. According to a home workout survey conducted in 2018, the recognition rate in association with "home workout" was high, more respondents were women in their 20s, and the proportion of home workouts was at a high level. Information related to home workouts was often obtained through YouTube or portal sites, and the major advantage was that users could exercise at any time and place they wanted without caring about others' eyes [22]. In contrast, the limited exercise methods, lack of motivation, and risk of misbehaving were indicated as disadvantages. Methods to obtain information about home s include online-based channels such as YouTube, Instagram, Internet search, applications, other online-based channels, acquaintances, and long-established knowledge [21].

Therefore, machine learning algorithms were applied to expand research methods in continuous participation factor analysis. We then analyzed the factors influencing the continuous exercise participation by the Keep users. Furthermore, we introduce the use of machine learning to determine whether to participate in continuous exercise centered on Keep.

2.2. Keep

Home workout apps have gained popularity with widely available mobile personal technology like smartphones and wristbands, years before the COVID-19 outbreak. These innovations focus on the consumer's market and correspond with self-tracking and self-management trends, promoting extensive engagement through interactive design elements. For those who want to exercise everywhere and keep track of their fitness status, fitness apps can easily help users be physically active anywhere and everywhere. The apps help manage their daily physical exercise, movement, or even dieting regiment, summarizing one's fitness status with quantified data charts such as the MyFitnessPal in the US or "Keep" in China [23].

Keep is an application dedicated to providing one-stop exercise solutions such as fitness teaching, running, cycling, making friends, fitness diet guidance, and equipment purchases. In February 2015, Keep 1.0 was officially released, and functions available mainly included training programs and fitness video training. The initial user size was 1 million. After the release of Keep 2.0 in August 2015, new interactive sharing features, including personal data and sharing of dynamics and global rankings, were developed. By early 2016, the number of users had grown 32% from August 2015, reaching 20 million. Released in April 2016, version 3.0 officially launched two functions: execution and e-commerce, marking the transition of the application from the first mobile fitness tool to an e-commerce sports platform exploring commercialization.

The application emphasizes "discipline" and "management"—exercise every day, monitor every action, share with friends and family, improve the overall level of activity, thereby obtaining a virtual upgrade of the user's status. These aspects are aimed at facilitating the participation of users in various functions of the application. For commercial applications, the goal is to expand the user base and increase engagement. Monthly active users (MAU) and daily active users (DAU) are the two most important indicators for evaluating the success of mobile applications and their profit potential [23]. Keep is available in 15 languages, including Chinese, Korean, Japanese, French, and German. Two related products, namely Keep Trainer and Keep Yoga, were launched on Google Play and App Store and operated on Instagram and Facebook [7]. This paper believes that the popularity of fitness applications occurs in the contextualized interaction between human and technical design.

2.3. Machine Learning Methodology

Machine learning is a method of implementing AI in which computers derive new information or make decisions by learning data using algorithms and programs, similar to human learning [6]. Conceptually, it is similar to data mining. Data mining refers to a set of tasks that can derive meaningful patterns, knowledge, or information from large amounts of data to produce results that can help make decisions [24].

The concept of AI was proposed in the 1950s, and it depends on neural network algorithms. In the 1980s, theoretical research on machine learning algorithms was considerably developed, but its application based on hardware development was quite limited. By default, the installation of machine learning algorithms is limited to a naïve Bayesian classifier, k-nearest neighbors (KNN), induction of rules and trees, support vector machines (SVM), neural networks, linear and logistic regression, and ensemble methods.

Machine learning can be useful when the size of the data is too large for humans to solve because it can be defined mathematically. However, the mathematical definition of complex problems is difficult [25]. Thus, results below the human level in terms of

real-world performance are often provided or are not available or difficult to be used in real-world environments.

Recently, the use of machine learning has rapidly expanded due to the activation of big data and the remarkable development of hardware, along with the development of deep learning algorithms that are more advanced than machine learning. Based on machine learning, patterns can be automatically extracted from data. Generally, machine learning can be divided into supervised learning, unsupervised learning, and reinforcement learning [26]. Algorithms in machine learning are usually used for prediction, classification, and clustering [27].

Unsupervised learning is a method that analyzes unanswered data [28] and finds clusters similar to input values in situations where users do not know to which class the data belong. The algorithm finds and produces similarities on its own, and this process enables the identification of the characteristics of each group [29]. Examples such as clustering of data, detection of anomalies, and speculation of data distribution are the most different to map learning [30], where the nature of the data is directly guessed with no need to create a function to calculate a particular value.

If unsupervised learning is utilized as a classification method, the model to be used in data prediction analysis is based on supervised learning. Map learning enables a model to automatically learn the relationship of attributes from past data [21]. Unsupervised learning exploits unanswered data, but in the case of supervised learning, the answer comes from the data, which creates a general decision rule [23].

Machine learning has been widely used in many fields, including disease diagnosis [31], stock trend prediction [32], image and speech recognition [33,34], and information extraction [35]. In the machine learning approach, a prediction model can be trained using input data to achieve a goal without solving theoretical equations.

2.3.1. Random Forest

Random Forest (RF), developed by Breiman, is a machine learning algorithm with many decision trees [36,37]. The input vectors for RF are represented as $\{x = [x_{i1}, x_{i2}, \cdots, x_{iM}], y_i\}$, $i = 1, 2, \cdots, m$, where M refers to the number of features which is expressed in Figure 1, m is the number of observations, and y is a scalar. The modeling of RF, which is expressed in Figure 2, consists of three parts: (1) Sample selection. The training data sets are extracted from the original data by using bootstrap P times. Bootstrap is a method in which each observation has the same chance of being selected in each sampling round. Therefore, some observations can be chosen more than once in P bootstrap rounds. After selecting samples, the training datasets for building the tree-based predictors are represented as $\{T_1, T_2, \cdots, T_p\}$. (2) Tree-based learners generation. The P learners (predictors) are generated by P training datasets. For each tree-based predictor, D features $(D < M)$ are randomly selected from M features for node splitting instead of using all features. That is, both the number of tree-based predictors and the number of selected features play important roles in the prediction performance of RF. (3) Result combination. Each tree-based predictor has the same contribution to the predicted result. For classification, the predicted result is the mode value of these tree-based predictors.

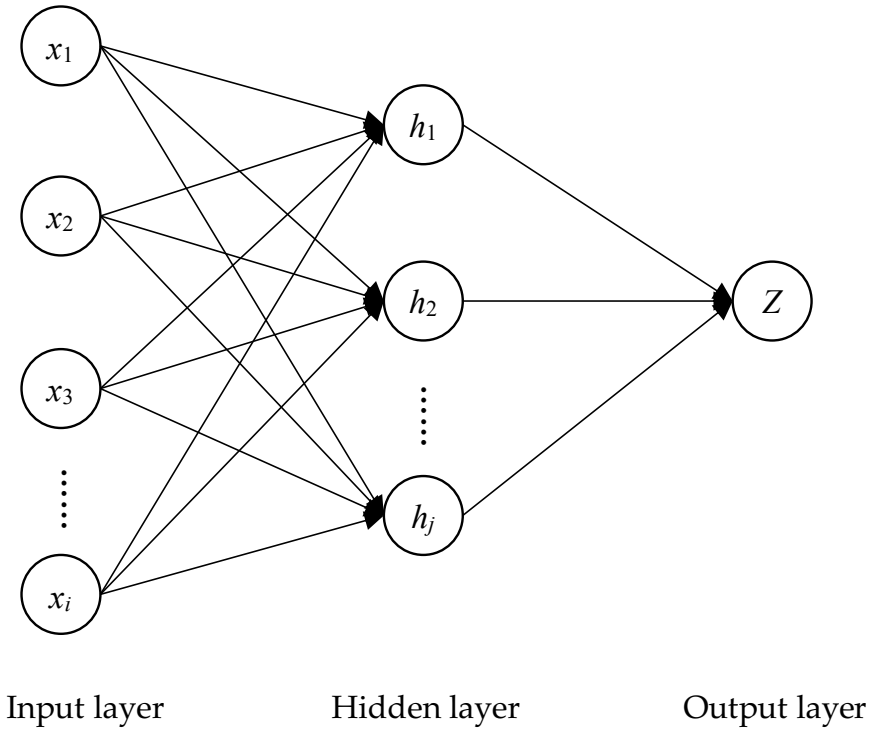


Figure 1. Structure example.

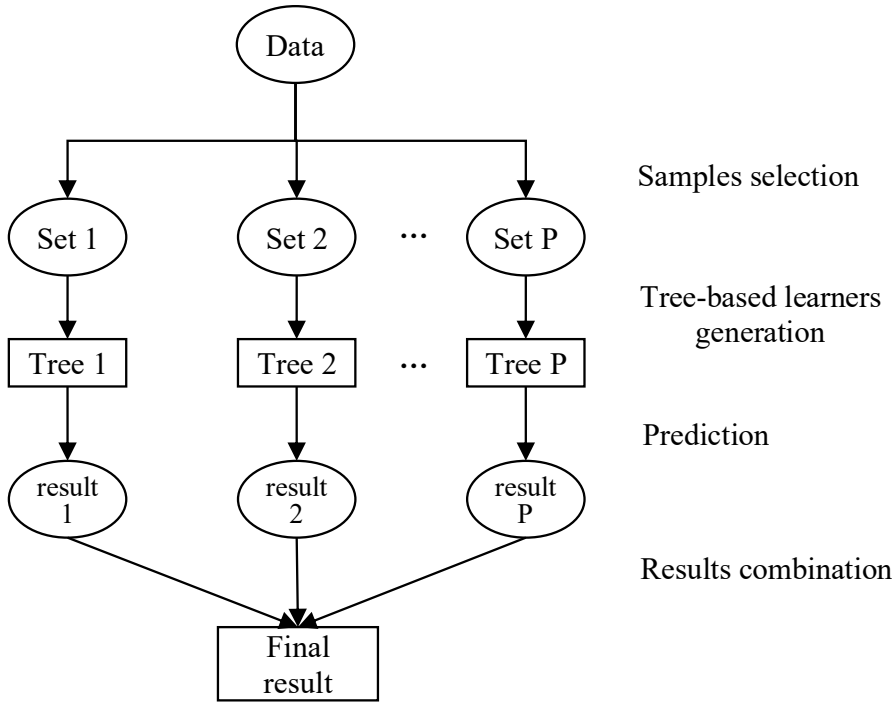


Figure 2. Modeling process of Random Forest.

The regression result is obtained by averaging the predicted results of these combined predictors.

2.3.2. Support Vector Machine

Support vector machine (SVM), proposed by Cortes and Vapnik, makes full use of the structural risk minimization theory (SRM), which ensures the strong generalization ability of the model [38]. Assuming that the observation i combines m-dimensional input

vectors $x_i \in R^m$ with a scalar output y_i, Equation (1) is defined to express the regression relationship between the nonlinear input and output in the SVM model:

$$f(x_i) = \omega^T \cdot \varnothing(x_i) + b,\ i = 1, 2, 3, \cdots, m \tag{1}$$

where ω and b are two parameters in the regression using SVM, the $f(x_i)$ represents the predicted results, the $\varnothing(x_i)$ is the mapping function that can transform the nonlinear vectors in a low-dimensional feature space to the linear vectors in a high-dimensional feature space.

To improve the generalization ability, the slack factors ζ_i and ζ_i^* are added to Equation (1) as they can reduce fitting errors. Then, the optimized regression function and its subjective conditions are expressed as Equation (2):

$$\begin{gathered} min\ R(\omega, \zeta_i, \zeta_i^*) = C\sum_{i=1}^{m}(\zeta_i + \zeta_i^*) + \tfrac{1}{2}\|\omega\|^2 \\ \text{s.t. } y_i - \omega^T\varphi(x_i) - b \le \varepsilon + \zeta_i \\ \omega\varphi(x_i) + b - y_i \le \varepsilon + \zeta_i^* \\ C > 0,\ \zeta_i \ge 0,\ \zeta_i^* \ge 0,\ \varepsilon \ge 0 \end{gathered} \tag{2}$$

where penalty C is used to obtain the balance between empirical risk and model complexity item $\|\omega\|$. Non-optimal C may lead to lower prediction accuracy and weaker generalization ability.

As a nonlinear optimization problem, Equation (2) can be solved by using duality theory based on the Lagrange multipliers method. Therefore, the original problem is transformed into the following dual problem:

$$\begin{gathered} \alpha^{(*)}\varepsilon R^{2m} minW\left(\alpha^{(*)}\right) = \tfrac{1}{2}\sum_{i,j=1}^{m}\left(\alpha_i^* - \alpha_i\right)\left(\alpha_j^* - \alpha_j\right)K(x_i, x_j) + \varepsilon\sum_{i=1}^{m}\left(\alpha_i^* - \alpha_i\right) - \sum_{i=1}^{m}\left(\alpha_i^* - \alpha_i\right) \\ \text{s.t. } \sum_{i=1}^{m}\left(\alpha_i - \alpha_i^*\right) = 0 \\ \alpha_i \ge 0,\ \alpha_i^* \le \tfrac{C}{m} \\ \sum_{i=1}^{m}\left(\alpha_i + \alpha_i^*\right) \le C \cdot v \end{gathered} \tag{3}$$

where $K(x_i, x_j)$ is the kernel function, such as a radial basis function (RBF), which must satisfy Mercer's theorem, and α_i, α_i^* are both Lagrange multipliers.

After solving the dual problem, all the parameters of SVM are determined. The new regression function of SVM is as follows:

$$f(x_i) = \sum_{i=1}^{m}(\alpha_i^* - \alpha_i)K(x_i, x_j) + b \tag{4}$$

2.3.3. K-Nearest Neighbor

K-nearest neighbor (KNN), proposed by Cover and Hart, is one of the most widely used predictive models. The prediction process is as follows: (1) Calculating the distance between the target and other known observations; (2) Extracting K neighbors that are close to the target in terms of the calculated distance; (3) Obtaining the value of the K neighbors; (4) Calculating the predicted value of the target. The regressive and classified results of the target are the average and mode value of the K neighbors, respectively [39]. Thus, the prediction of the target is affected by the number of neighbors (K) around the target and the selection of the distance function. Predicted results of the target based on a few neighbors are not sufficiently convincing, whereas excessive neighbors may contain noise that can lead to erroneous predictions. That is, a suitable K value is essential for the prediction

performance of KNN. In addition, the Euclidean distance function is defined as the physical distance between two observations [40], expressed as follows:

$$d_{ij} = sprt\left(\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2\right) \tag{5}$$

where d_{ij} represents the Euclidean distance between observations i and j. Observation i is written as $\{x_i = [x_{i1}, x_{i2}, x_{i3}, \cdots, x_{iM}], y_i\}$; observation j is represented as $\{x_j = [x_{j1}, x_{j2}, x_{j3}, \cdots, x_{jM}], y_j\}$.

2.3.4. Logistic Regression (LR)

Given a training set $\left\{\left(\vec{x}_1, y_1\right), \cdots, \left(\vec{x}_m, y_m\right)\right\}$, $\vec{x}_1 \in R^n$ and $y_i \in \{0,1\}$, $1 \le i \le m$, logistic regression (LR) defines the class of $\vec{x}_i$ by Equation (6):

$$y_i' = \begin{cases} 0, & p_i < 0.5 \\ 1, & p_i \ge 0.5 \end{cases} \tag{6}$$

where p_i is calculated using Equation (7) and σ is the sigmoid function defined by Equation (8):

$$p_i = \sigma\left(\vec{w}\vec{x}_i\right), \tag{7}$$

$$\sigma(a) = \frac{1}{1 + e^{-a}}. \tag{8}$$

The goal of the LR algorithm is to learn weights $\left(\vec{w}\right)$ that minimize a cost function, such as a cross-entropy function with ridge regression, which reduces the complexity of the model and prevents overfitting. The cross-entropy function with ridge regression is given by Equation (9).

$$J\left(\vec{w}\right) = -\sum_{i=1}^{m} y_i \log(p_i) + (1 - y_i)\log(1 - p_i) + \frac{\lambda}{2}\sum_{j=1}^{m} w_j^2, \tag{9}$$

where λ is a regularization parameter. The gradient descent algorithm starts with some initial $\vec{w}$ and repeats the update in Equation (10) until the termination criteria are satisfied.

$$\vec{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}_{n\times 1} + \alpha \begin{bmatrix} -\vec{x}_1- \\ -\vec{x}_2- \\ \vdots \\ -\vec{x}_m- \end{bmatrix}_{m\times n} \times \left(\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} - \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_m \end{bmatrix} \right)_{m\times 1} - \lambda \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}_{n\times 1} \tag{10}$$

where α is the learning rate, which affects how expeditiously the learning model can converge to local minima. If α is too small, slow convergence may occur; if α is too large, gradient descent can overshoot a local minimum. α and λ are important hyper-parameters of LR models. Tutun et al. applied the evolutionary strategy and simulated annealing to optimize the coefficients of LR [41].

3. Research Design

Many studies on machine learning techniques or data mining use the sample, explore, modify, model, assess (SEMMA) methodology. In the present study, we used the expanded SEMMA methodology to add the application phase and finally select the test samples for the existing SEMMA [42]. Following the SEMMA methodology, we performed data preprocessing based on the collected data, implemented a discriminant model based on the four machine learning algorithms, evaluated their performance, and applied test samples to the algorithm with the best discriminant verification.

3.1. Data Collection

In this study, we aimed to select independent variables that were expected to affect the determination of whether or not to exercise continuously through Keep, as well as determine whether users continuously participate in-home workouts. Thus, supervised learning was employed. Supervised learning is usually utilized to predict the relationship between independent variables and present a clear learning method for learning objects using data with objective and dependent variables (motion in this month or not) [43].

In the current work, Keep users were selected for analysis by considering Keep characteristics. In addition, the characteristics and related data of Keep users were collected. We utilized Web Crawler techniques (Python 3) to collect a total of 7734 data (training set) points and to obtain various variables related to the target variable, as shown in Figure 3. In addition, we publicly uploaded the code used in this study to GitHub [41]. Our public dataset can be accessed using the following URL: https://github.com/chenqiuying1023/keep/blob/master/keep_app_collector.py (accessed on 11 October 2021).

```python
def get_userinfo(user_id):
    url='https://api.gotokeep.com/social-user/v1/people/home?userId={}'.format(user_id)
    response=session.get(url=url)
    json=response.json()

    userBasicInfo=json['data']['headInfo']['userBasicInfo']
    username=userBasicInfo['username']
    gender=userBasicInfo['gender']
    province=userBasicInfo['location']['province']
    memberStatus=userBasicInfo['memberStatus']
    joinTime=userBasicInfo['joinTime']

    userSocialStaticsInfo=json['data']['headInfo']['userSocialStaticsInfo']
    # print(userSocialStaticsInfo)
    totalEntries=userSocialStaticsInfo['totalEntries']
    follow=userSocialStaticsInfo['follow']
    followed=userSocialStaticsInfo['followed']
    liked=userSocialStaticsInfo['liked']

    completeCount=json['data']['headInfo']['creatorStatsInfo']['completeCount']
    try:
        verifyDesc=json['data']['headInfo']['userVerifyInfo']['verifyDesc']
    except TypeError:
        verifyDesc=''
```

```python
}
response=session.get(url=url,headers=headers)
json=response.json()

userSportSummary=json['data']['userSportSummary']
minutesDuration=userSportSummary['minutesDuration']
currMonthMinutesDuration=userSportSummary['currMonthMinutesDuration']
timeStamp=userSportSummary['joinTime']
timeArray = time.localtime(timeStamp)
joinTime = time.strftime("%Y-%m-%d", timeArray)

sportSummaries_list=userSportSummary['sportSummaries']
sportSummaries_dict={}
for sportSummaries in sportSummaries_list:
    key=sportSummaries['trainingType']
    value=sportSummaries['formatTitleNumber']
    sportSummaries_dict[key]=value

try:
    training=sportSummaries_dict['training']
except KeyError:
    training=0
try:
    running=sportSummaries_dict['running']
```

Figure 3. Data crawler.

A total of 18 predictors were selected by analyzing the characteristics of Chinese Keep users. The variables collected are presented in Table 1.

3.2. Creating and Selecting a Discriminant Model

In the current work, we utilized supervised learning classification techniques to create discriminant models. Among the various supervised learning methods, the target variables were determined based on random forest, support vector machine, k-nearest neighbor, and logistic regression techniques. Orange (Ver. 3.28), which is one of the easiest-to-use data mining tools available, was used as an analysis tool for determination. To increase the reliability of the discriminant model, support vector machine, k-nearest neighbor, and the logistic regression technique used k-fold cross-validation, which was performed 10 times. Cross-validation [44–46], sometimes called rotation estimation [47–49] or out-of-sample testing, is any of various similar model validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. In a prediction problem, a model is usually given a dataset of known data on which training is run (training dataset) and a dataset of unknown data (or first seen data) against which the model is tested (called the validation dataset or testing set) [50,51]. The goal of cross-validation is to test the model's ability to predict new data not used in estimating it to flag problems like overfitting or selection bias [52] and to give insight on how the model will generalize to an independent dataset. We selected the model with the highest analytical index figures from the four above-mentioned models for continuous participation.

Table 1. Attributes of dataset.

Dataset		Explanation
Type/Characteristic	**Category**	
Basic Information	Gender	Male/female
	Sign-up time	Period of subscription to the Keep application
	Area	User's residential area (0: if no information is entered)
	VIP member status	Whether the user is a VIP member
User Level	T level	Total number of minutes of exercise from signing up to now (minutes)
	W level	Total health hours from signing up to now (minutes)
	R level	Total running distance (km) from entry until now
	C level	Total walking distance (km) from signing up to now
	Y level	Total yoga hours from signing up to now (minutes)
	B level	Total biker distance from signing up to now (km)
Data log	Number of badges	Number of badges acquired by the user
	Number of posts	Quantity of contents posted on Keep
	Number of followers	Number of users who like and follow an account, such as a specific person or business, on a social networking service
	Number of following	Number of users following
	Number of likes	Quantity of "likes" such as in posts
	Level	Keep level
	Quantity to follow within a month	Number of people who exercise according to the user per month
Keep certification	Keep Certification	When users use certain functions in Keep, they need to be authenticated first. Only after the identity certification, are they allowed to use this function. The user needs to provide ID card information that matches the person's identity, etc.
Sustainable use	Continuous participation in Keep	Participation in this month's exercise through Keep -1 for participation. -0 if not participating.

4. Evaluation and Selection of Discriminant Model

4.1. Discriminant Model Performance Validation Indicator

Among the implemented analysis methods, we employed the highest discriminant accuracy as a model-specific performance validation indicator in the present study to select the optimal analysis technique and used variable selection to validate the performance of the newly derived variables. Therefore, we adopted the error matrix of binary classification to calculate the reproducibility and precision of each model.

For the performance evaluation in the experiment, first, we denoted TP, FP, TN, and FN as true positives (the number of instances correctly predicted as required), false positive (the number of instances incorrectly predicted as required), true negative (the number of instances correctly predicted as not required) and false negative (the number of instances incorrectly predicted as not required), respectively. Then, we obtained four measurements: accuracy, precision, recall, and F1-measure, Table 2. True and false indicate whether the values determined were correct or not. In other words, true indicates that the discriminant value is equal to the actual value, whereas false indicates that the discriminant value is different from the actual value. Positive and negative indicate whether the determined values are positive or negative. Therefore, true positive (TP) indicates that the real value is positive and correctly determined. True negative (TN) means that the actual value is negative and correctly determined. In contrast, false positive (FP) suggests that the actual value is negative and determined to be positive, whereas false negative (FN) indicates that the actual value is positive but determined to be negative [53]. Based on these error matrices, we can create analytical metrics that can evaluate discriminant models.

Table 2. Numerical values of confusion matrix.

Contents	Meaning	Numerical Value
Accuracy	Ratio of the exact actual value to the actual value (positive to determine the actual voice by negative)	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	The percentage of positive values determined by positive values	$\frac{TP}{TP+FP}$
Recall	Percentage of positive determination of actual positive values	$\frac{TP}{TP+FN}$
F1-Score	Indicators evaluated by considering both factors using the harmonic mean of precision and Recall	$2 * \frac{precision*recall}{precision+recall}$

To enhance precision, we need to determine when the probability of the class will be extremely high. However, at this time, the performance of recall decreases, as part of the data that actually belong to the class is excluded due to low probability. Thus, the F1 score, which equally considers reproducibility and precision, can be useful in this case [37].

4.2. Receiver Operating Curve of a Discriminant Model

Based on the analysis metrics previously described, the discriminant power of the four models was evaluated. Thus, we performed a receiver operating curve (ROC) analysis. ROC is a plot of the FP (real voice classification as positive) ratio on the x-axis and the TP (real positive classification) ratio on the y-axis.

A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The method was originally developed for operators of military radar receivers starting in 1941, which led to its name.

The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, recall, or probability of detection [54] in machine learning. The false-positive rate is also known as probability of false alarm [48] and can be calculated as (1-Specificity).

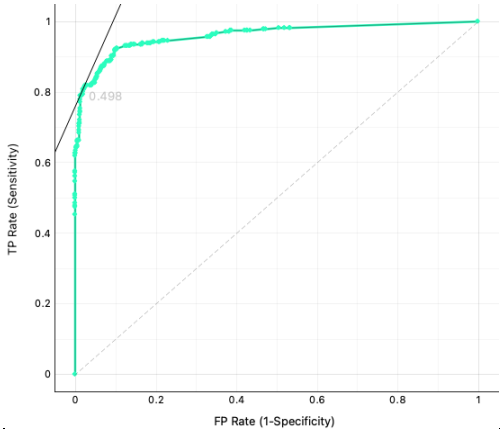
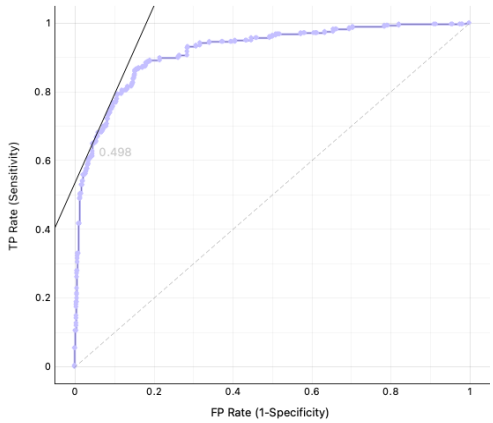
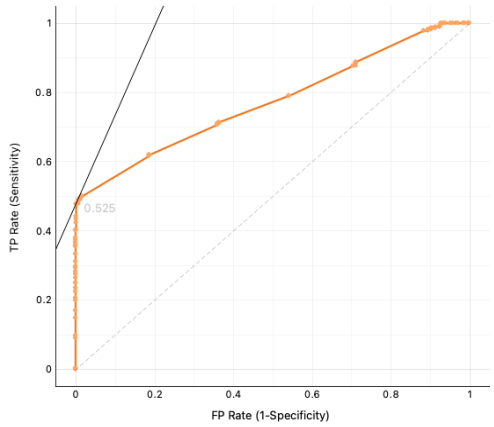
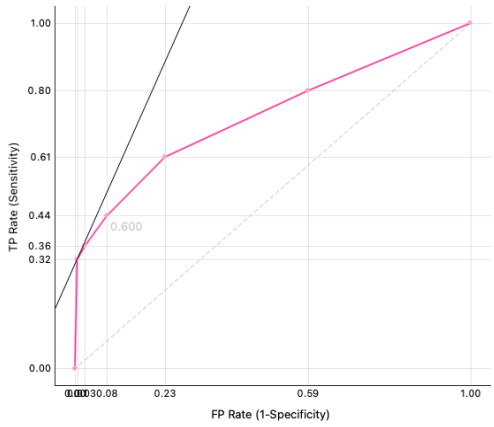
It can also be thought of as a plot of the power as a function of the Type I Error of the decision rule. The ROC curve is thus the sensitivity or recall as a function of fall-out.

Thus, when the curve is more convex upward, a higher TP ratio value and a lower FP ratio value denote better accuracy of the model. The lower part of this ROC curve is referred to as the area under the ROC curve (AUC) [55]. The value range of AUC is within 0.5 and 1. The ROC analyses are shown in Table 3. The AUC values for random forest reached the peak of 0.960, compared with 0.916 for logistic regression, 0.772 for SVM, and 0.734 for KNN. Furthermore, given the high precision, reproducibility, and F1 score of the random forest discrimination algorithm in the discrimination of the current work, the overall discrimination verification power of random forest was better than that of the other three models, as shown in Table 4.

Table 3. Predictability of evaluation model.

Model	AUC	F1	Precision	Recall
Random Forest	0.960	0.928	0.930	0.930
Logistic Regression	0.916	0.859	0.861	0.864
SVM	0.772	0.832	0.850	0.846
KNN	0.734	0.770	0.773	0.786

Table 4. ROC Curve.

Random Forest (AUC: 0.960)	Logistic Regression (AUC: 0.916)
	
Support Vector Machine (AUC: 0.772)	**KNN (AUC: 0.734)**
	

5. Evaluation of Test Sample Discrimination

5.1. Validation Dataset Discrimination

In order to avoid overfitting, when any classification parameter needs to be adjusted, it is necessary to have a validation data set in addition to the training and test datasets. The models analyzed can be applied to the actual test samples to reevaluate their effectiveness, consequently raising new problems and repeating previous tasks, if necessary. As described in the previous section, we discriminated test samples using the random forest algorithm, which was evaluated as the best discriminant verification model among the four discriminant models.

The validation Dataset is the sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration. A validation dataset consists of examples used to tune the hyperparameters of a classifier. An example of a hyperparameter for artificial neural networks includes the number of hidden units in each layer [56]. It, as well as the testing set, should follow the same probability distribution as the training data set. For further discussion, this study selected 10% of the training data as the validation set and again predicted the Keep result based on random forest.

Table 5 shows the binary classification result of the test samples. In the test samples, from 850 Keep users, 716 were determined to continuously use the application, and 15 were incorrectly identified as non-persistent users. Among the 119 test users who did not use Keep continuously, 87 were correctly determined and 32 were incorrectly determined. Using the random forest model, we verified this again; ROC analysis showed that the overall discriminant verification ability of random forest is better than the other three evaluation models. Table 6 shows the result was as follows: AUC = 0.88, precision = 0.979, reproducibility = 0.891, F1 = 932. As the training dataset decreased, it could be seen that the accuracy and precision were lower than the initial model. AUC is the area under the ROC curve, and its area will not be greater than 1. Since the ROC curve is generally above the straight line y = x, the value range of AUC is usually between (0.5, 1). Therefore, the verification of the random forest model is established.

Table 5. Confusion matrix for random forest.

		Predicted		Σ
		0	**1**	
Actual	0	87	32	119
	1	15	716	731
Σ		102	748	850

Table 6. Predictability of random forest.

Model	AUC	F1	Precision	Recall
Random Forest	0.988	0.932	0.979	0.891

5.2. Evaluation of Variables That Affect the Determination of Continuous Participation

After performing the learning process to determine continuous usage, we further analyzed the variables (18 variables in total) collected from a private homepage to identify more information and determine continuous participation based on the random forest model.

Figure 2 shows the rank of the contribution of variables that affected the prediction. Keep certification (0.374) was ranked first as the most important variable for continuous approval, followed by total exercise time (0.233), number of badges (0.230), and level (0.212). The remaining variables presented comparatively neglectable effectiveness in determining the continuous participation of respondents.

Keep certification means that among Keep users, someone who has verified their identity information can obtain the Keep certification logo. The identity information can only be obtained by providing real information verification such as ID card information, mobile phone verification, or email verification. Keep certification is the most important variable that affects the results in this study, indicating that the platform still needs to be improved in terms of real identity privacy information and other aspects.

6. Conclusions

In this study, machine learning technology was used to analyze the home workout field. We used machine learning algorithms to obtain factors that affect the continued use of a home workout platform. Simultaneously, we conducted a more scalable study of the analysis results to find real meaning and present effective solutions. The ROC analysis showed that the overall discrimination verification power of random forest was better than that of the other three models evaluated (AUC = 0.960, precision = 0.930, reproducibility = 0.930, F1 = 928). Thus, the proposed approach was helpful to predict the continuous usage of users after installing the application.

The results indicated that Keep certification appeared to be the most important cause for continuous exercise status which shows in Figure 4. Keep certification means that the user has verified their identity information. This shows that the platform still needs to be improved in terms of real identity privacy information and other aspects.

	#	Gain ratio	Gini
C KEEP certification	56	0.149	0.374
N T level		0.217	0.233
N Number of badges		0.214	0.230
N Level		0.239	0.212
N W level		0.179	0.187
N R level		0.136	0.145
N C level		0.113	0.113
N Y level		0.086	0.089
N B level		0.081	0.075
C Area	30	0.035	0.074
N Number of followers		0.066	0.073
N Number of following		0.044	0.049
C Gender	2	0.075	0.041
N Number of posts		0.037	0.040
N Quantity to follow within a month		0.038	0.021
N Likes		0.014	0.016
T Sign up time		0.012	0.013
C VIP Member Status	2	0.011	0.005

Figure 4. Ranking of predicted contributions of variables.

7. Results and Implications

We investigated the social distancing environment, which is currently a common topic, and examined the changes in a non-face-to-face online home workout service. As previously described, the random forest algorithm discrimination model obtained through machine learning provided accurate discrimination results in testing. Because Keep is China's largest home workout platform and manages a considerable amount of user information, it can utilize machine learning to determine whether its users are continuously using the application.

The academic implications of this study are as follows. First, we attempted to expand our study methods in the field of home workouts by applying machine learning. As home workout companies have been recently identified, related studies are few, despite the fact that the continuous use of users has become a problem faced by companies. With advances in AI technology, machine learning is efficient in processing information with new research methods and exerts an important role in helping to discriminate and process large amounts of data. In contrast to traditional social science analysis methods, we chose a machine learning approach to make predictions. Machine learning can solve problems faster and more easily, reveal insights, and thus indicate operating plans for similar home workout platform companies and provide competitive advantages.

Second, it provided indications about the factors influencing continuous exercise participation. We studied the influence of each of the 18 variables that were expected to affect the determination of continuous participation. Among them, the variables that could be reflected had minimal influences, such as the sign-up time, VIP membership status, and the number of likes received. On the contrary, we found that certification, which refers to the verification of the user's identity through peer review, is the most critical factor for continuous use. This indicates that providing factual information verification, such as ID card information, mobile phone verification, or email, highly motivates users to continue working out. The reason is that as the amount of exercise increases, status certification and trust are needed. In this regard, it is believed that the continuous participation of users will increase if companies provide easier methods to gain certification or trust. For home workout companies, machine learning provides operational solutions, such as how to enhance personal certification aiming to increase the continuous use of the platform.

Third, due to personal privacy reasons, we only collect public data from users, so we cannot collect more personal information. Since its launch in 2015, Keep has more than 300 million users and over 10 million members. However, according to the results, users who did not continue using Keep accounted for a large proportion of all test sets, which was different from the usage rate found in market research. We need to collect more data from more platforms to investigate user adherence to home workout platforms.

Author Contributions: Conceptualization, S.L.; methodology, S.L.; software, Q.C.; validation, Q.C. and S.L.; data curation, Q.C.; writing—original draft preparation, Q.C.; writing—review and editing, S.L.; visualization, Q.C.; supervision, S.L.; project administration, S.L.; funding acquisition, S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by supported by the BK21 FOUR Program, funded by the Ministry of Education (MOE, Korea) and the National Research Foundation of Korea (NRF).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This work was supported by an Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2019-0-01343).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Grenita, H.; Deepika, R.L.; Shane, A.P.; Carl, J.L.; Ross, A. A tale of two pandemics: How will COVID-19 and global trends in physical inactivity and sedentary behavior affect one another? *Public Health Emerg. Collect.* **2021**, *64*, 108–110. [CrossRef]
2. Jong, S.T.; Drummond, M.J.N. Exploring online fitness culture and young females. *Digit. Leis. Cult.* **2016**, *35*, 758–770. [CrossRef]
3. Mutz, M.; Müller, J.; Reimers, A.K. Use of digital media for home-based sports activities during the COVID-19 pandemic: Results from the German SPOVID survey. *Int. J. Environ. Res. Public Health* **2021**, *18*, 4409. [CrossRef]
4. WHO. Healthy at Home—Physical Activity. Available online: https://www.who.int/news-room/campaigns/connecting-the-world-to-combat-coronavirus/healthyathome/healthyathome---physical-activity (accessed on 12 January 2021).
5. Thompson, W.R. Worldwide survey of fitness trends for 2021. *ACSMs Health Fit. J.* **2021**, *25*, 10–19. Available online: https://journals.lww.com/acsm-healthfitness/Fulltext/2021/01000/Worldwide_Survey_of_Fitness_Trends_for_2021.6.aspx (accessed on 22 January 2021). [CrossRef]
6. Panch, T.; Szolovits, P.; Atun, R. Artificial intelligence, machine learning and health systems. *J. Glob. Health* **2018**, *8*, 020303. [CrossRef] [PubMed]
7. Zhou, K. Keep Becomes China's Largest Social Sports Platform. Available online: https://pandaily.com/Keep-becomes-chinas-largest-social-sports-platform/ (accessed on 20 August 2017).
8. Dunn, M.R.; DeJonckheere, M.; Schuiteman, S.; Strome, A.; Herbert, K.; Waselewski, M.; Chang, T. "Stay home so this can be over:" A national study of youth perspectives on social distancing during the COVID-19 pandemic. *Prev. Med. Rep.* **2021**, *22*, 101355. [CrossRef]
9. Kim, R.; Jeon, M.; Lee, H.; Choi, J.; Lee, J.; Kim, S.; Lee, S.; Seo, Y.; Kwon, J. *Trend Korea 2018*; Publishing Co.: Seoul, Korea, 2018.
10. Chaabene, H.; Prieske, O.; Herz, M.; Moran, J.; Höhne, J.; Kliegl, R.; Ramirez-Campillo, R.; Behm, D.G.; Hortobágyi, T.; Granacher, U. Home-based exercise programmes improve physical fitness of healthy older adults: A PRISMA-compliant systematic review and meta-analysis with relevance for COVID-19. *Ageing Res. Rev.* **2021**, *67*, 101265. [CrossRef]
11. Ahmad, A.; Garhwal, S.; Ray, S.K.; Kumar, G.; Malebary, S.; Barukab, O.M. The Number of Confirmed Cases of Covid-19 by using Machine Learning: Methods and Challenges. *Arch. Comput. Method Eng.* **2020**, *28*, 1–9. [CrossRef] [PubMed]
12. Felice, F.D.; Polimeni, A. Coronavirus disease (COVID-19): A machine learning bibliometric analysis. *In Vivo* **2020**, *34*, 1613–1617. [CrossRef]
13. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2008.
14. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
15. McCall, B. Covid-19 and artifcial intelligence: Protecting health-care workers and curbing the spread. *Lancet Digit. Health* **2020**, *2*, e166–e167. [CrossRef]
16. Pham, Q.V.; Nguyen, D.C.; Huynh-The, T.; Hwang, W.J.; Pathirana, P.N. Artificial Intelligence (AI) and big data for coronavirus (COVID-19) pandemic: A survey on the state-of-the-arts. *IEEE Access* **2020**, *8*, 130820–130839. [CrossRef]
17. Naudé, W. *Artifcial Intelligence Against COVID-19: An Early Review*; RWTH Aachen University: Aachen, Germany; IZA: Bonn, Germany, 2020. Available online: https://www.iza.org/publications/dp/13110/artificial-intelligence-against-covid-19-an-early-reviewBatis (accessed on 6 April 2020).
18. Bullock, J.; Luccioni, A.; Pham, K.H.; Lam, C.S.N.; Luengo-Oroz, M. Mapping the landscape of artificial intelligence applications against COVID-19. *J. Artif. Intell. Res.* **2020**, *69*, 807–845. Available online: https://arxiv.org/abs/2003.11336v3 (accessed on 19 November 2019). [CrossRef]
19. Vaishya, R.; Javaid, M.; Khan, I.H.; Haleem, A. Artifcial intelligence (AI) applications for covid-19 pandemic. *Diabet. Metab. Syn. Clin. Res. Rev.* **2020**, *14*, 337–339. [CrossRef]
20. Sá-Caputo, D.D.C.D.; Taiar, R.; Seixas, A.; Sanudo, B.; Sonza, A.; Bernardo-Filho, M. A proposal of physical performance tests adapted as home workout options during the COVID-19 pandemic. *Appl. Sci.* **2020**, *10*, 4755. [CrossRef]
21. Wire, Business. Worldwide IT Training Market Trends, Share, Size, Growth, Opportunity and Forecast (2019–2024). Available online: https://www.businesswire.com/news/home/20190412005400/en/Worldwide-Training-Market-Trends-Share-Size-Growth (accessed on 12 April 2019).
22. Jung-hee, O.; Jae-woo, O.; Kwang-min, C. Research on consistent use intention of home-training program on personal media service Youtube based on post-adoption model. *J. Conv. Soc. Korea* **2019**, *10*, 183–193. [CrossRef]
23. Zheng, E.L. Interpreting fitness: Self-tracking with fitness apps through a postphenomenology lens. *AI Soc.* **2021**, 1–12. [CrossRef]
24. Shmueli, G.; Bruce, P.C.; Yahav, I.; Patel, N.R.; Lichtendahl, K.C., Jr. *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*; Wiley: Hoboken, NJ, USA, 2017.
25. Moon, S.; Jang, S.B.; Lee, J.H.; Lee, J.S. Trends in machine learning and deep learning technologies. *Inf. Commun. Mag.* **2016**, *33*, 49–56.
26. Kelleher, J.D.; Namee, B.M.; D'Arcy, A. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*; MIT Press: Cambridge, MA, USA, 2020.
27. Kouroua, K.; Exarchos, T.P.; Exarchos, K.P.; Karamouzis, M.V.; Fotiadis, D.I. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* **2015**, *13*, 8–17. [CrossRef] [PubMed]
28. Nakai, Y.; Takiguchi, T.; Matsui, G.; Yamaoka, N.; Takada, S. Detecting abnormal word utterances in children with autism spectrum disorders: Machine-learning-based voice analysis versus speech therapists. *Percept. Mot. Ski.* **2017**, *12*, 961–973. [CrossRef] [PubMed]

29. Marsland, S. *Machine Learning: An Algorithmic Perspective*, 2nd ed.; Chapman & Hall/CRC Press: Los Angeles, CA, USA, 2014.
30. Schmidt, J.; Marques, M.R.G.; Botti, S.; Marques, M.A.L. Recent advances and applications of machine learning in solid-state materials science. *npj Comput. Mater.* **2019**, *5*, 83. [CrossRef]
31. Soni, J.; Ansari, U.; Sharma, D.; Soni, S. Predictive data mining for medical diagnosis: An overview of heart disease prediction. *Int. J. Comput. Appl.* **2011**, *17*, 43–48. [CrossRef]
32. Patel, J.; Shah, S.; Thakkar, P.; Kotecha, K. Predicting stock and stock price index movement using trend deterministic datapreparation and machine learning techniques. *Expert Syst. Appl.* **2015**, *42*, 259–268. [CrossRef]
33. Ciregan, D.; Meier, U.; Schmidhuber, J. Multi-column deep neural networks for image classification. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3642–3649.
34. Hinton, G.; Deng, L.; Yu, D.; Dahl, G.; Mohamed, A.R.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T.; et al. DeepNeural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.* **2012**, *29*, 82–97. [CrossRef]
35. Freitag, D. Information extraction from HTML: Application of a general machine learning approach. In Proceedings of the AAAI-98 Proceedings, American Association for Artificial Intelligence, Palo Alto, CA, USA, 26–30 July 1998; pp. 517–523.
36. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
37. Iranitalab, A.; Khattak, A. Comparison of four statistical and machine learning methods for crash severity prediction. *Accid. Anal. Prev.* **2017**, *108*, 27–36. [CrossRef]
38. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
39. Cover, T.M.; Hart, P.E. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [CrossRef]
40. Hewson, P.J. Multivariate Statistics with R. Available online: https://www.docin.com/p-445285843.html (accessed on 20 July 2012).
41. Tutun, S.; Khanmohammadi, S.; He, L.; Chou, C.A. *A Meta-Heuristic LASSO Model for Diabetic Readmission Predictio*; Industrial & Systems Engineering Research Conference (ISERC): Anaheim, CA, USA, 2016.
42. Barrios, M.; Jimeno, M.; Villalba, P.; Navarro, E. Novel data mining methodology for healthcare applied to a new model to diagnose metabolic syndrome without a blood test. *Diagnostics* **2019**, *9*, 192. [CrossRef]
43. Gwak, J.; Yoon, H.S. Development of a model for winner prediction in TV audition program using machine learning method: Focusing on program. *Knowl. Manag. Res. Pract.* **2019**, *20*, 155–171. [CrossRef]
44. Allen, D.M. The relationship between variable selection and data agumentation and a method for prediction. *Technometrics* **1974**, *16*, 125–127. [CrossRef]
45. Stone, M. Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. Ser. B (Methodol.)* **1974**, *36*, 111–147. [CrossRef]
46. Stone, M. An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *J. R. Stat. Soc. Ser. B (Methodol.)* **1977**, *39*, 44–47. [CrossRef]
47. Geisser, S. *Predictive Inference*; Chapman and Hall: New York, NY, USA, 1993.
48. Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, Montreal, QC, Canada, 20–25 August 1995; pp. 1137–1143.
49. Devijver, P.A.; Kittler, J. *Pattern Recognition: A Statistical Approach*; Prentice-Hall: London, UK, 1982.
50. Galkin, A. What Is the Difference between Test Set and Validation Set? Available online: https://stats.stackexchange.com/questions/19048/what-is-the-difference-between-test-set-and-validation-set (accessed on 10 October 2018).
51. Newbie Question: Confused about Train, Validation and Test Data! Available online: https://en.wikipedia.org/wiki/Cross-validation_(statistics) (accessed on 14 November 2016).
52. Cawley, G.C.; Talbot, N.L.C. On over-fitting in model selection and subsequent selection bias in performance evaluation (PDF). *J. Mach. Learn. Res.* **2010**, *11*, 2079–2107. [CrossRef]
53. Müller, A.C.; Guido, S. *Introduction to Machine Learning with Python A Guide for Data Scientists*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2016.
54. Detector Performance Analysis Using ROC Curves—MATLAB & Simulink Example. Available online: www.mathworks.com (accessed on 11 August 2016).
55. Keller, H.; Müller, L.M.; Schraner, T.; Kellenberger, C.J.; Saurenmann, R.K. Is early TMJ involvement in children with juvenile idiopathic arthritis clinically detectable? Clinical examination of the TMJ in comparison with contrast enhanced MRI in patients with juvenile idiopathic arthritis. *Pediatr. Rheumatol.* **2015**, *13*, 56. [CrossRef]
56. Ripley, B.D. *Pattern Recognition and Neural Networks*; Cambridge University Press: Cambridge, UK, 1996; p. 354.