Machine Learning based Prediction of The Value of Buildings

Woosik Lee[1], Namgi Kim[2], Yoon-Ho Choi[3], Yong Soo Kim[4], and Byoung-Dai Lee[2]*
[1]Research Center, Social Security Information Service, Seoul, South Korea
[2]Department of Computer Science, Kyonggi University, Suwon, South Korea
[3]Department of Computer Science, Pusan National University, Busan, South Korea
[4]Department of Industrial and Management Engineering, Kyonggi University, Suwon, South Korea
[e-mail: blee@kgu.ac.kr]
*[1]Corresponding author: Byoung-Dai Lee

Received Feburary 26, 2018; revised May 14, 2018; accepted June 30, 2018;
published August 31, 2018

Abstract

Due to the lack of visualization services and organic combinations between public and private buildings data, the usability of the basic map has remained low. To address this issue, this paper reports on a solution that organically combines public and private data while providing visualization services to general users. For this purpose, factors that can affect building prices first were examined in order to define the related data attributes. To extract the relevant data attributes, this paper presents a method of acquiring public information data and real estate-related information, as provided by private real estate portal sites. The paper also proposes a pretreatment process required for intelligent machine learning. This report goes on to suggest an intelligent machine learning algorithm that predicts buildings' value pricing and future value by using big data regarding buildings' spatial information, as acquired from a database containing building value attributes. The algorithm's availability was tested by establishing a prototype targeting pilot areas, including Suwon, Anyang, and Gunpo in South Korea. Finally, a prototype visualization solution was developed in order to allow general users to effectively use buildings' value ranking and value pricing, as predicted by intelligent machine learning.

Keywords: Machine Learning, Random Forest, Fully Connected Network, Deep Learning

This work was supported by the GRRC program of Gyeonggi province.
A preliminary version of this paper was presented at ICONI 2017, and was selected as an outstanding paper.

http://doi.org/10.3837/tiis.2018.08.023 **ISSN : 1976-7277**

1. Introduction

National organizations, local governments, and public institutions recently have opened data to the public. This data had accumulated in a variety of fields, including education, health and medical services, public administration, social welfare, and science and technology. This development has ignited the creation of new businesses and jobs that make use of the data. In particular, the Basic Map of National Land Information, maintained and managed by the National Research Institute, contains a variety of serial cadastral maps and ortho-images. Through integration with multiple spatial information, the basic map can be applied to diverse convergent industrial fields, such as ubiquitous, mobile, and virtual spaces. A task remains in maximizing the utility of the Basic Map of National Land Information for space industries. Based on this need, mid- and long-term development roadmaps should be established in order to create new value and to reinforce the competitiveness of relevant industries.

However, there have been very few studies investigating the organic combination between public and private data or creating diverse services aimed at maximizing the usability of the Basic Map of National Land Information. Virtually no studies have attempted to create new services by combining the new technology of big data with the machine learning technique.

This study establishes a database by combining public and private data that currently are separated, thus developing a solution that provides visualization services to general users. To this ends, an in-depth study was implemented addressing three major aspects. First, factors that could potentially affect the building price were examined and relevant data attributes defined. This paper thus presents a data acquisition method for extracting relevant data attributes, according to information sources. Such data include public spatial information provided by the central and local governments and real estate-related information provided by the private real estate portal sites. The paper also introduces the pretreatment process required for intelligent machine learning. Second, this report suggests that intelligent machine learning can predict value pricing and the future value of buildings; to do so, it relies on big data regarding buildings' spatial information, as obtained from a database containing building value attributes. The availability of the algorithm was tested by establishing a prototype using Suwon, Anyang, and Sanbon in South Korea as a pilot area. Third, a prototype visualization solution was developed to support the general users' effective use of buildings' value ranking and value pricing, as predicted by intelligent machine learning.

Therefore, the purpose of this paper implement machine learning based prediction program and visualization prototype using public and private combined data set.

This paper will proceed in the following manner. Chapter 2 explains domestic and international trends regarding real estate information systems in the public sector and machine learning algorithms. Chapter 3 examines schema for the establishment of databases in terms of buildings' value attributes and the data acquisition method. Chapter 4 investigates representative machine learning algorithms that predict building values. Chapter 5 analyzes the experimental results regarding the effectiveness of building value prediction. Chapter 6 introduces a visualization solution that supports effective use by general users. The final chapter summarizes the study and presents future study plans.

2. Related Works

This chapter examines representative domestic and international public information systems and reviews existing studies related to machine learning-based prediction of real estate value.

In South Korea, several public portal systems contain real estate information, including the Onnara Integrated Portal of Real Estate Information, [1] the real Estate Transaction Management System (RTMS) [2], and Naver Real Estate [6].

The Onnara Integrated Portal of Real Estate Information [1] is the world's first real estate information portal serviced by a government. The service provides information about the price of nationwide real estate, including land and houses (the actual traded price of apartments, appraised value of land, declared price of house, and market value of apartments), as well as information about land use regulations by lot using map visualization. It also provides real estate-related civil services so that users can print confirmation documents regarding land use plans or register real estate online without visiting government offices. RTMS [2] is an integrated portal-based report system containing the actual traded price, established to secure transparency and trade order in the real estate market. Naver Real Estate (land.naver.com) [6] is a representative real estate portal service in Korea that provides real estate sales information as well as information about the surrounding environment, including cadastral maps, aerial photographs, and street views. In addition to these systems, the Korean Appraisal Board [3] provides statistical data about real estate, the Korean Credit Bureau [4] supports the establishment of regional public policies and real estate while providing credit information, and Wisenet [5] offers a counseling system for real estate customers. Finally, Jik Bang is an information system specializing in annual and monthly rents [7], and LOBIG provides market value information by using public data and artificial intelligence [8].

Similar to the real estate information system in Korea, many public information systems about real estate also have been established in the United States, the United Kingdom, and Japan.

In the United States, information related to the real estate market is produced and managed by the National Statistical Office, the Federal Housing Finance Agency, the Department of Housing and Urban Development (HUD), and private organizations [9]. In the United Kingdom, real estate-related statistics [10] are produced and managed by the National Statistical Office, Land Registry, the Department for Communities and Local Government (DCLG 5), and private organizations. Interform by DCLG 5 is a system that compiles data from the local governments. Interform distributes Excel files with survey items along with a manual for writing each statistics item. The system retrieves the completed files and compiles the data. The UK Land Registry announces monthly market trend data, including the House Price Index, transaction data, and recent sales price information. In Japan, real estate-related statistics are produced and managed by the Statistics Bureau at the Ministry of Internal Affairs and Communications; the Ministry of Land, Infrastructure, Transport and Tourism; the Japan Real Estate Research Institute; the Real Estate Distribution System; and the Tokyo Stock Exchange [11]. The Japan Real Estate Research Institute and the Real Estate Distribution System handle statistics related to real estate prices and transactions. The Statistics Bureau handles information related to housing supply and current issues in the real estate market. The Ministry of Land, Infrastructure, Transport, and Tourism has taken action to improve the real estate transaction systems, creating the Real Estate Information Center, which collaborates with the Real Estate Information Network System (REINS). The Real Estate Information Center is a system that provides comprehensive real estate information combining transaction information, public information, and regional and population information through collaboration with the REINS. Transaction information includes price, ownership history, guarantees, construction costs, appraisal evaluations, and diagnostic reports. Public information includes taxes, land use zoning, history of flooding, history of maintenance, and real estate maps. Along with this data, the system also provides population and regional information.

Although no previous study has specifically analyzed national public data based on intelligent machines, several studies have considered the vision of using a national public information platform or machine learning [12]-[17].

Kang (2016) [12] studied a method for connecting and integrating an open platform of spatial information with a national public information database. The study investigated the technical schemes required to collaborate with V World, which is an open platform spatial information service provided by the Ministry of Land, Infrastructure, and Transport. Unfortunately, the study followed a rather simple survey format in terms of the national public information database and did not include plans for specific database use or applying development systems. Sim (2016) [13] introduced a method for determining real estate values by using machine learning algorithms, classifying artificial intelligence into four phases. Specifically, the study explained the following four phases in the value determination process for machine learning: choosing the real estate for value determination, data establishment and pretreatment, algorithm exploration and modeling, and practical operation and sophistication process. However, the study failed to apply the presented scheme to a specific case. Cho (2016) [14] described the usability and outlook of a large data-based artificial intelligence system in the real estate field. The study introduced the current situation in the real estate service industry by summarizing information about the service based on theme classification and regional classification. The study also examined the classification system of spatial information and the number of datasets in each category. The study elaborated on the map learning algorithm using an actual example. Unfortunately, the study did not explain how to apply a specific machine learning scheme. Chung [15] conducted research into predicting the price index of multi-unit houses using an artificial neural network. The study implemented an experiment by directly building an artificial neural network composed of 12 intelligence neural networks. Due to the simplicity of the artificial neural network model, the researcher recommended more specific experiments applying up-to-date techniques. Yeon (2015) [16] employed a study using a logistic regression model and decision-making tree in order to improve the accuracy of homes' standard declared price based on the machine learning-based approach. The study adapted an ensemble model using bagging and gradient boosting. However, it was an early study into the declared price of houses, and it lacked sufficient investigation into advanced model applications. Lee and Park (2016) [17] examined applications of the machine learning model for estimating the price of detached houses. Through a detailed comparison analysis, the paper examined non-parametric models' characteristics and pros and cons, including GAM, RF, MARS, and SVM. The study also conducted an actual experiment evaluating model performance. However, the study implemented no specific analysis of the deep learning scheme. Hwang (2017) [18] presented diverse cases related to combining big data in the real estate market. However, the paper faces a limitation: it did not cover the use of artificial intelligence schemes or specific realization methods. In our research, we first implemented an integration of public and private data before predicting future value based on the established database referring to a machine learning algorithm. This paper also discusses a solution for prototype visualization that assists effective use by general users.

3. Model for Transmission-Power Control

This chapter examines potential factors that affect building prices before defining the related data attributes. To extract the relevant data attributes, the data acquisition method is explained according to the information source, including the public spatial information data provided by the central and local governments and real estate-related information provided by the private real estate portal sites.

3.1 Database Schema

This section examines the potential factors affecting building price and defines the schema for establishing a database. Fig. 1 illustrates the data schema used to establish a building database. This information is based on apartments' physical attributes, including accessibility, surrounding environment, and block attributes of 0.5 kilometers, 1.0 kilometers, and 1.5 kilometers.

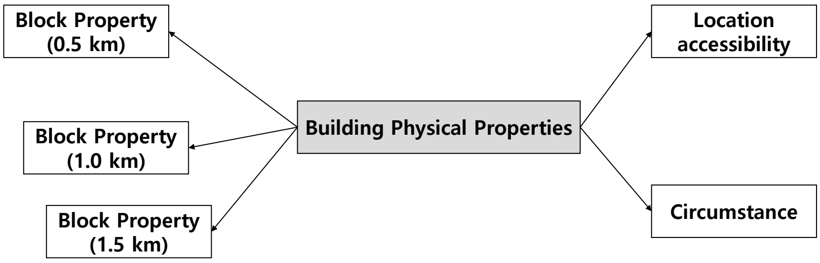


Fig. 1. Structure of Building Physical Properties

The physical attributes of apartment includes the following: pyeong (approximately 3.3 square meter) type of each building, as that is distinguished by the GIS building integration identification number;, price per pyeong;, legal dong code;, legal dong name; of dong, lot number; building, name; of the building uilding, sub- name; of the building, address by street name code;, eup-myeon-dong-ri and basement code of address by street name;, main and sub-house number of the address by street name;, upper level building identification number;, the number of ground stories;, date of approval;, date of user permission;, ratio of floor area to site;, level of deterioration;, total number of households;, number of households of each pyeong type of the complex;, brand value score;, walking time to elementary school;, and latitude and longitude location information.

Table 1 shows the column definition of the apartments' the physical attributes of the apartment. The item name is separated in English and Korean. Sample data shows the data that can be referred to when implementing the actual data collection. P0 is the GIS building integration identification number corresponding to that each building has. P1, which refers to pyeong type, refers to means the pyeong information of the apartment to the KB market value, with a where the unit value of is 10,000 KRW. P3 is a legal dong code, which is a ten-digit administrative district code indicating where the land is located. Legal dong refers to means the name of the administrative district where the land is located. P5-P25 references means specific information about of each building. P26 shows the score by brand of each apartment. P28 and P29 indicates means the latitude and longitude value of each building, respectively.

Table 1. Column definition of the physical attributes of apartments

Item name (English)	Item name (Korean)	Sample data
P0	Integrated GIS building ID No.	19952004360945276952000000000
P1	Pyeong type (Target)	66.12
P2	Price per pyeong (Label)	12300
P3	Legal dong code	4117310200
P4	Legal dong name	Gwanyang-dong, Dongan-gu, Anyang-si, Gyeonggi-do,
P5	Lot No.	37-3
P6-1	Building name (spatial information)	Ace Apt.
P6-2	Building name (Naver)	Ace Apt.
P7	Building No.	No. 101
P8	Road name address code	111104100096
P9	Road name address eup myeon dong ri code	1
P10	Road name address underground code	0
P11	Road name address main No.	00253
P12	Road name address sub No.	00007
P13	Parent building identification No.	100204360
P14	No. of stories	16
P15	Permit date	20160822
P16	Date of approval of use	20160822
P17	floor area ratio	67
P18	Deterioration level	17
P19	Total No. of households	1235
P20	No. of households by floor space (20 pyeong: less than 12m^2)	25
P21	No. of households by floor space (30 pyeong: less than 99.17m^2)	25
P22	No. of households by floor space (40 pyeong: less than 32.23m^2)	25
P23	No. of households by floor space (50 pyeong: less than 165.29m^2)	25
P24	No. of households by floor space (60 pyeong: less than 198.35m^2)	25
P25	No. of households by floor space (60 pyeong: more than 198.35m^2)	25
P26	Brand	10
P27	Walking time to primary school	6
P28	Location information (latitude)	176884.5095
P29	Location information (longitude)	431582.1237

Accessibility is an indicator that measures the accessibility of each building. It includes travel time to Gangnam, distance to highway interchange, distance to subway station, number of bus stops within a 200-meter range, and distance to bus stop.

Table 2 shows the column definition of accessibility. The item name is separated in English and Korean. Sample data show the information that can be referred to when implementing actual data collection. L0 refers to the unique GIS building integration identification number of each building. L1 indicates how long it takes to reach the Gangnam region for each building. L2 and L3 indicates the lineal distance from each building to the nearest highway IC and the nearest subway station, respectively. L4 refers to the number of bus stops within 200 meters of each building. L5 indicates the distance to the nearest bus stop from each building.

Table 2. Column definition of the location accessibility

Item name (English)	Item name (Korean)	Sample data
L0	Integrated GIS building ID No.	1995200436094527695200000000
L1	Entry time to Gangnam	125.12
L2	Distance to Expressway IC	125.12
L3	Distance to subway station	125.12
L4	No. of bus stops within 200m	7
L5	Distance to the bus stop	120.12

The "surrounding environment" refers to the existence of neighboring buildings around each building. The surrounding environment attributes include the distance from each building to a general hospital, large retailers, department stores, parks, public offices, McDonald’s, Starbucks, rental apartments, crematoriums, detention centers/prisons, brothels, and food waste composting facilities.

Table 3 shows the column definition for the surrounding environment. The item name is separated in English and Korean. Sample data show the information that can be referred to when implementing actual data collection. N0 refers to the unique GIS building integration identification number for each building. N1-N12 indicates the distance from each building to a general hospital, large retailers, department stores, parks, public offices, McDonald’s, Starbucks, rental apartments, crematoriums, detention centers/prisons, brothels, and food waste composting facilities. The distance between the building and each surrounding environmental feature was measured in meters.

Table 3. Column definition of the surrounding environment

Item name (English)	Item name (Korean)	Sample data
N0	Integrated GIS building ID No.	1995200436094527695200000000
N1	Distance to general hospital	1520
N2	Distance to supermarket	1520
N3	Distance to department store	1520
N4	Distance to park	1520
N5	Distance to public office	1520
N6	Distance to McDonald's	1520
N7	Distance to Starbucks	1520
N8	Distance to rental apartment	1520
N9	Distance to crematorium	1520
N10	Distance to detention center/prison	1520
N11	Distance to brothel	1520
N12	Distance to food recycle facility	1520

"Block attributes" include the average level of deterioration, the housing population figure, the average age of the housing population, district distribution information, the number of private educational institutes, and the number of middle and high schools within 0.5, 1.0, and 2.0 kilometers of each building.

Table 4 shows the column definition for block attributes. The item name is separated in English and Korean. Sample data show the information that can be referred to when implementing actual data collection.

Table 4. Column definition of the block attributes (0.5, 1.0, and 2.0km)

Item name (English)	Item name (Korean)	Sample data
BA0, BB0, BC0	Integrated GIS building ID No.	1995200436094527695200000000
BA1, BB1, BC1	Average deterioration level	12.3
BA2, BB2, BC2	Number of residential population	252
BA3, BB3, BC3	Mean age of the residential population	54.3
BA4, BB4, BC4	Land use zoning distribution 1	58.3
BA5, BB5, BC5	Land use zoning distribution 2	58.3
BA6, BB6, BC6	Land use zoning distribution 3	58.3
BA7, BB7, BC7	Land use zoning distribution 4	58.3
BA8, BB8, BC8	Land use zoning distribution 5	58.3
BA9, BB9, BC9	No. of private education institutes	52
BA10, BB10, BC10	No. of middle and high schools	52

The column definition of block attributes BA, BB, and BC was attached in front of the item name, according to the distance value of each block attribute. Here, the distance value of the block attribute has a range in a rectangular shape, with a center point indicating the location of each building separated by the GIS building integration identification number, as shown in **Fig. 2**. BA, BB, and BC refers to distance values of 0.5 kilometers, 1.0 kilometers, and 2.0 kilometers, respectively.

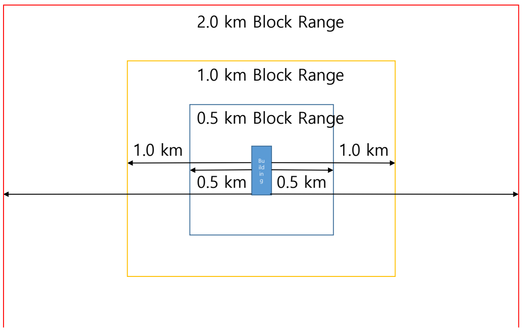


Fig. 2. Diagram of the block attribute range

Referring to average deterioration level, BA0, BB0, and BC0 represent the sum average of the deterioration level of each building in the relevant block. The size of the housing population refers to the number of people living in the relevant block. The average age of the housing population is the sum of the age of all people living in the relevant block divided by the number of people in the housing population.

Use district distribution is classified into residential zone, commercial zone, industrial zone, green belt zone, and management zone. Use district distribution 1 refers to a residential zone, and use district distribution 2 indicates a commercial zone. Use district distribution 3 references an industrial zone, and use district distribution 4 indicates a green belt zone. Finally, use district distribution 5 refers to a management zone. Residential zones include type 1 exclusively residential zones, type 2 exclusively residential zones, type 1 general residential zones, type 2 general residential zones, type 3 general residential zones, and semi-residential zones. Commercial zones include central commercial zones, general commercial zones, neighboring commercial zones, and retail commercial zones. Industrial zones include exclusively industrial zones, general industrial zones, and semi-industrial zones. Green belt zones include preservation green belt zones, production green belt zones, and natural green regions. Management zones include preservation management zones, production management zones, and program management zones.

3.2 Data Gathering Mechanism

This section examines the data acquisition method according to the information sources. These sources include public spatial information data provided by the central and local governments and real estate-related information provided by private real estate portal sites.

To extract the public spatial information data provided by the central and local governments, we located basic information using an open application programmer interface (API) provided by the Portal of Public Information Data and Data Center. Based on this basic information, specific information was extracted using the ArchGIS tool [21].

First, UQA information corresponding to building keywords was extracted from the public data. We approached the data using an SQL query statement in the ArcGIS program to extract UQA information. The first step was to recall shp files containing building polygons in ArcGIS. Once the shp file was recalled, a polygon was drawn on the ArcGIS main screen according to the information about each polygon. When a map is entered into the ArcGIS background, users can see which part of the current map corresponds to each polygon at a glance, allowing efficient selection of the desired polygon. The extraction is facilitated by the SQL query statement, which is widely used in the database.

To extract the desired data from the polygon existing in ArcGIS, the AWL query statement related to the data extraction should be entered into the window for SQL query. For example, to extract data containing the word "UQA122" in the A1 column, we must use an SQL query statement such as "A1" LIKE "%UQA122."

When the data are extracted by the SQL query statement, the color of the selected polygon changes so users can check the data expressed in different colors on the map. **Fig. 3** shows the blue color of a selected polygon in ArcGIS.

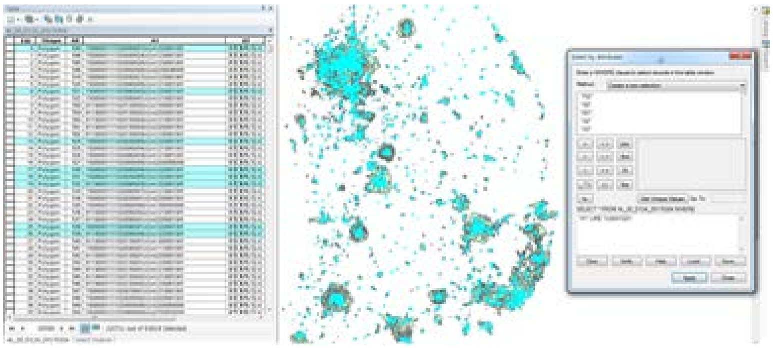


Fig. 3. The example of ArcGIS for region selections

The EDS Viewer Program allows extraction of latitude and longitude values in the left bottom and right top of each polygon box [22]. Based on these extracted latitude and longitude values, users can obtain the ratio that the relevant polygon occupies.

To obtain information about the bus stops located within 200 meters of a certain building, we used the service API of bus stop lookups in the public data portal. In particular, we determined the service API of bus stop lookups for the X and Y coordinates using the request variable shown in **Table 5**.

Table 5. Station Stop Service Request Variable

Item (Korean)	Item (English)	Item size	Item category	Sample data	Item description
X coordinate	x	10	Mandatory	127.1	X coordinate (WGS84)
Y coordinate	y	10	Mandatory	37.03	Y coordinate (WGS84)

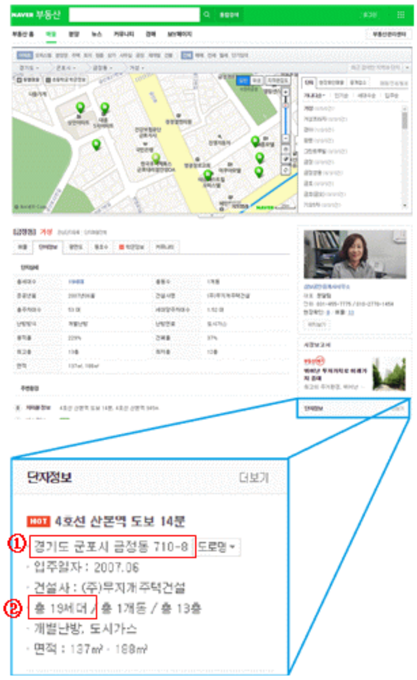
Table 6 shows the printed output received at this point in the process.

Table 6. Request results from the bus stop search service

Item (Korean)	Item (English)	Item size	Item category	Sample data	Item description
List of bus stops	BusStationAroundList	166	Mandatory		List of bus stops
Bus stop ID	stationId	166	Mandatory	231001214	Bus stop ID
Bus stop name	stationName	100	Mandatory	Pangok Middle School	Bus stop name
Bus stop No.	mobileNo	5	Mandatory	01234	Unique five-digit mobile No.
Region name	regionName	30	Option	Seongnam-si	Region name where the bus stop is
Control region	districtCd	1	Mandatory	1	Control region of the route
Median Bus Lane available?	centerYN	1	Option	Y	N: General Y: Median bus lane
X coordinate	X	10	Mandatory	127.109	X coordinate of bus stop
Y coordinate	Y	10	Mandatory	37.03	Y coordinate of bus stop

Different from the API provided by the central and local governments, extracting private information data requires morpheme extraction. Thus, we approached websites that provide private data and extracted data using morpheme analysis. This research established a database using information provided on the Naver Real Estate and KB Real Estate websites.

To extract essential data from the Naver Real Estate Service website, we first accessed the service and navigated to the needed screen. This screen showed the relevant complex after selecting the provincial units of do, si, gu, and dong, as well as the name of the complex. The new screen then provided diverse information about the complex, including occupancy date, construction company, and area. Here, we used the legal name of dong, lot number, and total number of households in the complex as the screen information. Users can locate information about the pyeong type of the relevant complex by selecting a floor plan type on the complex screen. Pyeong type information includes supplied area, area of exclusive use, and the number of households living in each pyeong type in the complex.



단지정보 더보기

HOT 4호선 산본역 도보 14분

① 경기도 군포시 금정동 710-8 도로명

- 입주일자 : 2007.06
- 건설사 : (주)무지개주택건설

② 총 19세대 / 총 1개동 / 총 13층

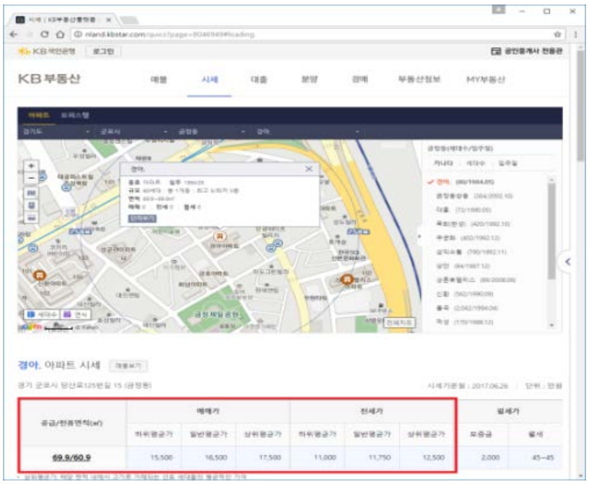
- 개별난방, 도시가스
- 면적 : 137㎡ - 168㎡

Fig. 4. Naver Real Estate Service Website

We next sought the market value of each complex from the KB Real Estate Service by the KB Kookmin Bank. The first step was to check each do, si, gu, dong, and complex on the main page and the market value page of the KB Real Estate Service. After clicking the market value in the top menu of the main page, the list of metropolitan cities, including Seoul and dos, appear on the right side along with the national map. Users can see the list of gu or cities in the region when they click on the desired region. Similarly, a list of dong pops up when users select gu or si. Finally, a list of complexes is presented after choosing the name of a dong.

Once the desired complex is chosen, the page displays sales price per each area, the lower average for rent prices, the general average for rent prices, and the upper average for rent prices (Fig. 5). Price information is updated every month, and users can see the past market value history by clicking the 'past market value lookup' button on the graph. Then, a window for past market value lookup pops up; here, the bottom of the page displays the sales price as well as lower, general, and upper average rent prices for the chosen month. The baseline month for the currently collected market value data is March 2017.

Different from the information method using parsing like the Naver Real Estate Service, the KB Real Estate Service has automated the collection of market value through a crawling scheme. The Selenium Webdriver module [19] was added to the previously-used scheme. After opening the past market value lookup page in the webdriver's phantomJS browser [20] and retrieving region and complex selections using the find_element_by_xpath() function, the contents can be changed using the click() function. Using this process, we are able to search and store in Excel the sales price and rent price of each pyeong type in the complex as of March 2017.



경아. 아파트 시세

경기 군포시 당산로125번길 15 (금정동)

시세기준일 : 2017.06.26 | 단위 : 만원

공급/전용면적(㎡)	매매가			전세가			월세가	
	하위평균가	일반평균가	상위평균가	하위평균가	일반평균가	상위평균가	보증금	월세
69.9/60.9	15,500	16,500	17,500	11,000	11,750	12,500	2,000	45~45

Fig. 5. KB Real Estate Service Website

4. Intelligence Machine Learning Algorithm

This chapter suggests an intelligent machine learning algorithm that predicts the value pricing and future value of buildings by using the big data of buildings' spatial information, as acquired from the building value attributes database. This chapter also tests the availability of the algorithm by establishing a prototype using the test area of Suwon, Anyang, and Sanbon. This paper utilizes Random Forest (RF) and Deep Neural Network (DNN) to suggest an intelligent machine learning algorithm.

4.1 Random Forest

This section examines the RF scheme used to predict buildings' value pricing and future value. The RF algorithm used for building value analysis in this study is a type of ensemble learning scheme for machine learning; it often is used for categorization and regression analysis. As shown in **Fig. 6**, the result class on the input data is printed from the multiple decision trees that were composed in the training process.

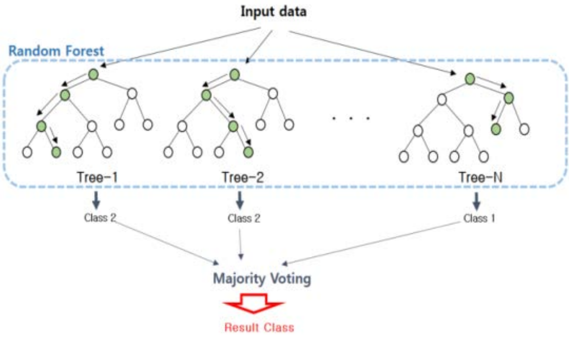


Fig. 6. Random Forest Operation Overview

The RF algorithm consists of a ‘learning phase’ that constitutes multiple decision trees and a ‘test phase’ that classifies the input vector and predicts the outcome. A forest composed of N number of decision trees is formed through the learning phase. When the test data come in, the final outcome is drawn through the voting of an outcome in each tree.

RF has the following competitive advantages compared to other classification algorithms: high accuracy, fast learning speed and testing, ability to process multiple variables without variable deletion, excellent generalization performance through randomization, and multiple class algorithm attributes.

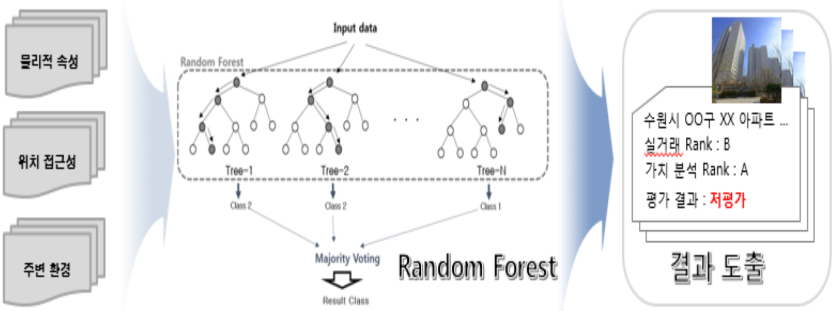


Fig. 7. Random Forest-based Value Analysis Overview

Fig. 7 provides a conceptual map of the value analysis. For the value analysis of buildings, we obtain the RF using rank, which is based on well-known building information (physical attributes, accessibility, and surrounding environment information) as well as price per pyeong. We begin by entering a certain building’s information requiring value analysis. After the learning process, the rank of the relevant building is drawn, and the users can determine whether the relevant building is overvalued or undervalued. The user makes this determination by comparing the actual rank and the test outcome rank.

RF-based building value analysis involves a total of six phases. Fig. 8 illustrates the flow for processing a value analysis.

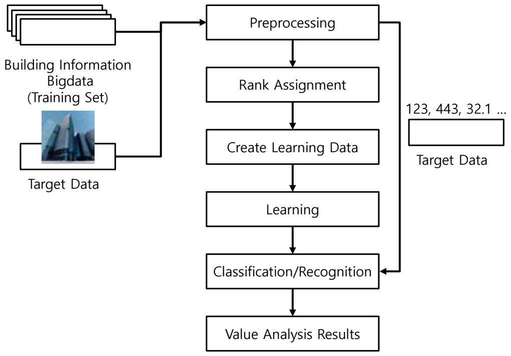


Fig. 8. Random Forest-based Value Analysis Operation Flowchart

A textual explanation of each phase follows below.

In the pretreatment phase, building information (attributes) is extracted for the RF learning from the spatial information data (training set) used for the learning process. **Table 7** summarizes the building information used for the RF learning.

Table 7. Attribute information for RF learning

Category of value attribute information	Use value attribute information
Physical attribute (14)	Pyeong type, road name address underground code, No. of stories, floor area ratio, deterioration level, total number of households, brand, walking time to primary school, and No. of households per pyeong in the complex (5)
Location accessibility (5)	Entry time to Gangnam (downtown), walking time to subway station, No. of bus stops, distance to bus stop, accessibility to IC
surrounding environment (10)	Distances to general hospital, supermarket, department store, public office, McDonald's, Starbucks, rental apartment, crematorium, detention center/prison, brothel, and food recycle facility
Value analysis criteria (1)	Rank based on price per pyeong

In the value rank assignment phase, value rank is assigned based on the price per pyeong of each building. The value rank is used as Class in RF, which becomes a criterion that determines the outcome of the value analysis.

In the learning data creation phase, learning data for the RF constitution is created by using building information extracted in the pretreatment phase and the relevant building's rank.

In the learning phase, the RF is created by using the learning data created from the previous phase. The RF learning outcome is presented below.

RF is composed of N number of trees designated by the user, and each decision tree is created by using 70% of the learning data, randomly selected from the entire learning dataset.

In the classification/recognition phase, value rank is drawn by using the data of a building whose value will be evaluated as RF input. The data used for the test is composed identically to the learning data. The final result of the value evaluation can be obtained by comparing the actual rank of the building and the one drawn from the RF.

The value analysis outcome phase visualizes the outcome extracted in the classification/recognition phase and provides it to the service users.

RF implementation can be divided into two steps: main flow and sub flow. The “main flow” refers to the whole forest building process and the “sub flow” refers to the process of creating one decision tree.

In the main flow, the data subset used for each tree creation is randomly selected, and tree creation is repeated until satisfying the RF learning termination condition (the number of trees). In the tree creation phase, the brunch of tree proceeds in the direction that has the highest information gain among the attributes constituting the data subset. **Fig. 9** shows an overall flow chart of the RF implementation described above.

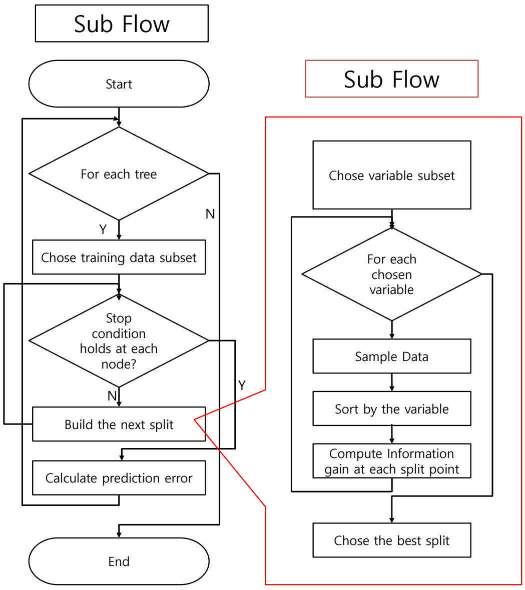


Fig. 9. Configuring the Random Forest Source Code Functionality

4.2 Deep Neural Network (DNN)

This chapter examines a DNN scheme for predicting buildings' value pricing and future value. The DNN algorithm used for the building value analysis is an inference algorithm in machine learning that uses a neural network structure composed of an input layer, output layer, and hidden layer. DNN is characterized by neurons that exist on each layer and are connected to all neurons on the output layer.

The DNN algorithm can be composed in three phases: the network constitution phase, the training phase, and the inference phase. The network composition phase determines the number of classes that constitute the whole network and the number of neurons in each class. The training phase searches for the values of optimal inference by updating the weight value that exists between each class. Finally, the inference phase draws the most optimal outcome by using the model developed in the training phase.

As shown in **Fig. 10**, DNN implementation begins by calling on the setting data to establish DNN. DNN setting data can determine the path, maximum epoch, learning rate, batch size, loss type, data composition ratio, and maximum fold. The DNN network is established based on this setting data. The composition differs according to the number of network classes and the number of neurons for each class at the time of network establishment. Once the network establishment is completed, training data is extracted. The composition depends on the data composition ratio of the setting data at the time of training data extraction. For example, if the data composition ratio is 20%, training data becomes 80% and test data becomes 20%. Implementation of training and inference is repeated up to the maximum fold according to the number of folds.

Fig. 11 summarizes the learning procedure of DNN. In the first phase, a part of the training data is randomly brought and trained in a mini-batch. The slope is computed in the next phase, with the slope of each weight mediation variable being obtained for the purpose of reducing the loss function value. The loss function is used as an index indicating the 'poorness' of the neural network's performance. In the third phase, the mediation variables are updated. For the

DNN of building value attributes, we updated the mediation variables by using an Adam scheme, which combines momentum and AdaGrad. It holds an advantage in that it supports efficient exploration of the mediation variable space and corrects hyper parameter bias.

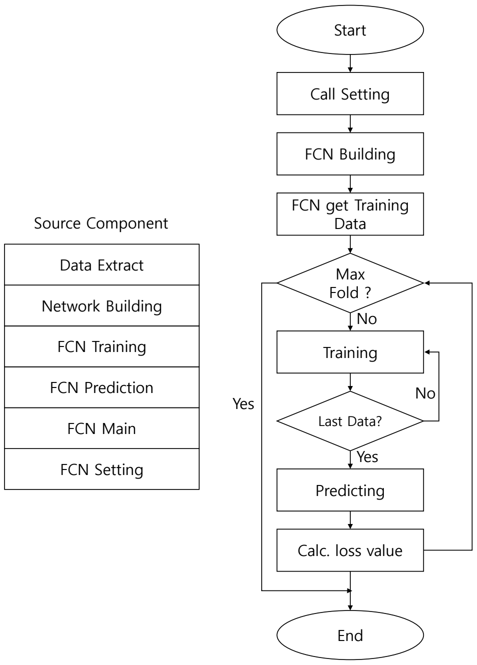


Fig. 10. Components and flows of FCN implementations

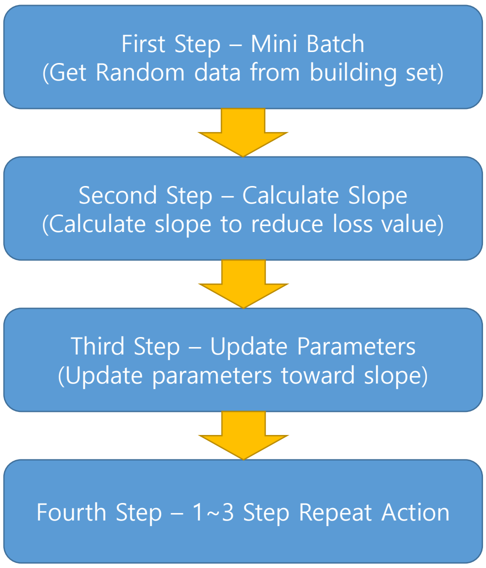

First Step – Mini Batch
(Get Random data from building set)
Second Step – Calculate Slope
(Calculate slope to reduce loss value)
Third Step – Update Parameters
(Update parameters toward slope)
Fourth Step – 1~3 Step Repeat Action

Fig. 11. Step 4 where FCN is performed

The level of movement is determined by obtaining the slope W for finding the F(x) in the building value analysis. As shown in Fig. 12, the smaller the W value, the longer it takes to access to the optimal point. On the contrary, a large W value makes approaching the optimal point faster, but it has the disadvantage of low accuracy, a factor called the ‘learning rate’ in DNN. In general, the most appropriate learning rate is chosen following diverse experiments. We began by manually ascertaining that 0.0005 is an appropriate value. Then, we searched for the optimal learning rate through an experiment using parameters between 0.00001-0.0009.

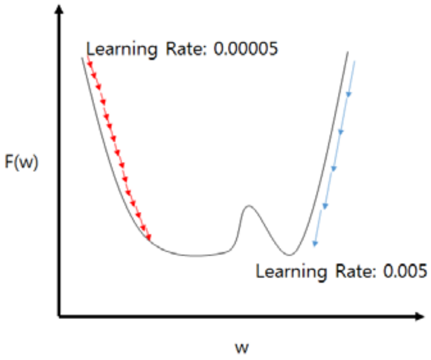


Fig. 12. Example of w Value Access by Learning Rate

The large size of the data from Suwon, Anyang, and Gunpo made possible implementation of training by cutting the data into mini batch sizes.

To analyze DNN outcomes in this paper, we proceeded with the experiment by creating divisions of two types, class and cost. Class consists of 0-7 phases and Cost refers to the actual measured price of the relevant building. Here, we implemented class analysis after classification according to the actual value and predicted value. For this research, we developed three categories: match, high, and low. These categories are detailed below.

1) Match: when the actual value is equal to the predicted value.
2) High: when the actual value is higher than the predicted value.
3) Low: when the actual value is lower than the predicted value.

As shown in **Fig. 13**, we analyzed accuracy by applying the models developed in Suwon, Gunpo, and Anyang through cross layer application.

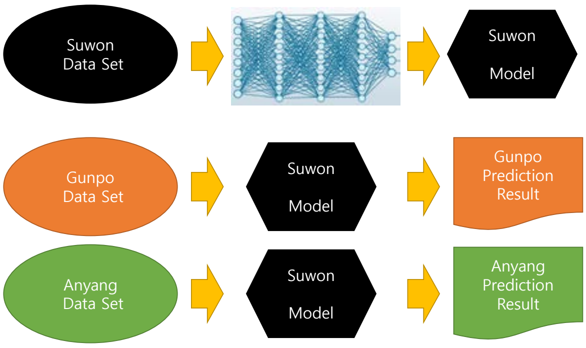


Fig. 13. Example of using Cross Layer to predict

5. Experiment

This chapter establishes a prototype targeting the test area of Suwon, Anyang, and Gunpo. The chapter then evaluates its availability using the intelligent machine learning algorithm presented in Chapter 4, which predicts buildings' value pricing and future value.

5.1 Experimental Environment

This section describes the experimental environment for predicting buildings' value pricing and future value.

In cases with no building data, we used the following scheme to prevent the loss of training data. If there existed value determination data for a pyeong type in the same building within a 10% range of the pyeong type, we used the same value determination data to fill in the missing data. If not, we altered the building data such that the values became zero. The training data was composed after excluding data with a zero value for the building. **Table 8** shows the total number of real estate datasets.

Table 8. Number of the modified real estate data sets

	Suwon	Gunpo	Anyang
No. of original data	46191	16101	22860
No. of space and filled data	12450	2031	3222
No. of zero	4491	1254	1875
Final file	41700	14847	20985

For DNN, the network was established by dividing it into three regions: Suwon, Anyang, and Gunpo. The layer of each network was in a fully-connected form where all input nodes were connected to all output nodes. Here, the dataset was trained using 20% of the sample data and 80% of the training dataset. DNN requires the optimal learning rate, epoch, and network layer numbers. Hence, it was necessary to search for appropriate training data parameters using experiments based on diverse scenarios. We first manually searched for the parameters broadly before specifically searching for specific parameters with script files. Finally, we conducted an experiment by applying the training data to find the optimal parameter values.

Table 9 shows the feature set for creating the training dataset for the DNN-based building value analysis.

Table 9. Column definition of the FCN training data

Item name (English)	Item name (Korean)	Item name (English)	Item name (Korean)
P1	Pyeong type	L3	Distance to subway station
P10	Road name address underground code	L4	No. of bus stops within 200m
P14	No. of stories	L5	Distance to the bus stop
P17	floor area ratio	N1	Distance to general hospital
P18	Deterioration level	N2	Distance to supermarket
P19	Total No. of households	N3	Distance to department store
P20	No. of households of 20 pyeong	N4	Distance to park

P21	No. of households of 30 pyeong	N5	Distance to public office
P22	No. of households of 40 pyeong	N6	Distance to McDonald's
P23	No. of households of 50 pyeong	N7	Distance to Starbucks
P24	No. of households of 60 pyeong	N8	Distance to rental apartment
P25	No. of households of 60 or larger pyeong	N9	Distance to crematorium
P26	Latitude	N10	Distance to detention center
P27	Longitude	N12	Distance to food recycle facility
L1	Entry time to Gangnam	Rank	Brand value of the building
L2	Distance to Expressway IC	Cost	Building price

As shown in Fig. 14, we separated the training data column such that the training set became 80% and the test set became 20%.

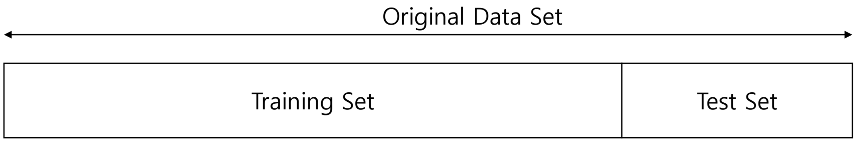


Fig. 14. Schematic diagram of the training data

We partitioned the test into five to test all the data that exist in the original dataset. There was no overlapping data for any test set.

The original dataset in one-hot encoding form was set as false, storing the values from 0 to n. One-hot encoding form refers to a sequence that has 0 for all the elements except for the elements that indicate the correct answer. For example, the sequence [0,0,0,0,1] has 0 for all elements except for one. If the one-hot encoding is false, the label in numerical values other than 1, such as ‘3’ or ‘2,’ can be stored. Fig. 15 shows the dataset that is partitioned into five.

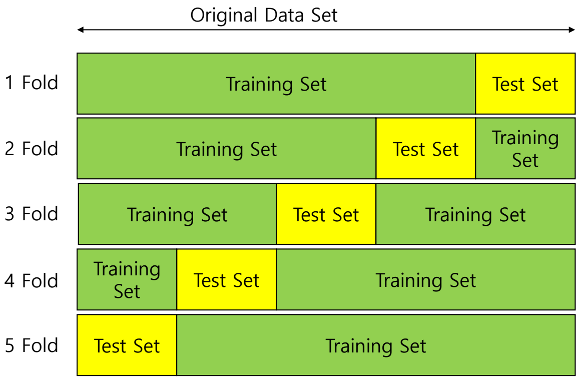


Fig. 15. Data set consisting of 5 for training data

By applying each dataset to the model, we created the data file with columns shown in Table 10. We made the building address by combining the items in the P4, P6-1, and P7 columns. The latitude and longitude of each building also was extracted from the P28 and P29 columns. We present the actual price or brand value of each building in the next column, and the final column shows the inference value obtained by applying the dataset to the model.

Table 10. Column definition of the results after data set applied to the model

Item name (English)	Item name (Korean)
P4	Legal dong name
P6-1	Building name (spatial information)
P7	Building No.
P28	Location information (latitude)
P29	Location information (longitude)
Real Cost or Rank	Actual building brand or Building price
Inference Cost or Rank	Building brand or building price inferred from the model

5.2 Experimental Analysis

To predict buildings' value pricing and future value, this section tests the availability in a pilot region, including Suwon, Anyang, and Gunpo, based on the experimental environment established in Section 5.1. Here, the RF and DNN machine learning algorithms explained in Chapter 4 are used for the availability test. The experiment proceeded by distinguishing between the same region and the different region.

Table 11 shows the results of the same region availability test, which analyzes the match between the actual value and the predicted value; in this case, it is the same in Suwon-Suwon, Gunpo-Gunpo, and Anyang-Anyang regions. The experimental outcome shows that RF predicts approximately 5-7% more accurately compared to DNN overall. The difference is largest in the case of predictions in Gunpo. Both algorithms demonstrated the most accurate predictions in Suwon.

Table 11. Verification of Same Region Validity (Match)

	Suwon	Gunpo	Anyang
RF	90.72%	82.52%	80.11%
DNN	86.55%	75.58%	75.89%

Table 12 shows the same region availability test, in the case of high, where the actual value is higher than the predicted value. It indicates that the Anyang region has a higher predicted value than other regions overall. The DNN algorithm evaluates the Suwon region higher than the Gunpo region, which differs from the RF algorithm.

Table 12. Verification of Same Region Validity (High)

	Suwon	Gunpo	Anyang
RF	3.57%	5.17%	7.95%
DNN	6.99%	5.46%	10.39%

Table 13 shows the same region availability test in case of low, where the actual value is lower than the predicted value. Both the RF and DNN give the lowest prediction values in the Gunpo region, which differs from the Suwon region. Similar to the high case, compared to DNN overall, RF demonstrates fewer instances in the low case where the predicted value is lower than the actual value.

Table 13. Verification of Same Region Validity (Low)

	Suwon	Gunpo	Anyang
RF	5.71%	12.31%	11.94%
DNN	6.46%	18.96%	13.72%

Table 14 shows the match case, where the actual value equals the predicted value in Gunpo and Anyang regions by using the Suwon model. Both the RF and DNN evaluate building values in Anyang higher than in Gunpo. RF has superior prediction rates for Gunpo, while DNN provides better predictions for Anyang.

Table 14. Verification of Different Region Validity (Suwon Model)

	Gunpo	Anyang
RF	19.67%	32.32%
DNN	17.39%	33.88%

Table 15 shows the match cases in the Suwon and Gunpo regions by using the Anyang model. Different from the Suwon model, the DNN algorithm provides better prediction rates than RF in the case of the Anyang model.

Table 15. Verification of Different Region Validity (Anyang Model)

	Suwon	Gunpo
RF	31.26%	23.42%
DNN	44.65%	34.70%

Table 16 shows the match cases in the Suwon and Anyang regions by using the Gunpo model. The Gunpo model shows a completely different outcome than the Anyang model. Overall, the RF experimental outcome shows more accurate prediction rates than the DNN experimental outcome. The difference in accuracy is approximately 17% in the case of Suwon.

Table 16. Verification of Different Region Validity (Gunpo Model)

	Suwon	Anyang
RF	30.36%	21.76%
DNN	13.92%	19.08%

After testing availability in the same region, the experimental results above imply that the RF scheme delivers approximately 4.5% better performance than the DNN scheme overall. When applied in different regions, both the RF and DNN schemes show a maximum performance below 50%. In terms of the performance, DNN performed better in the case of the Anyang model and RF was superior in the case of the Gunpo model.

Through the experimental outcome, we confirmed that the machine learning model can predict building values with an accuracy of over 80%. We also were able to demonstrate that the superior algorithm differs depending on the region.

Therefore, machine learning based building value prediction is realistic, and it is expected that better results will be obtained if various tuning and data refinement are added a little more.

7. Visualization Prototype

This chapter presents a solution for visualizing the prototype so that general users can effectively use buildings' value rankings and value pricing, as predicted by the intelligent machine learning.

Our visualization service for building value analysis includes 'search by value' and 'search by difference (difference of the value).' The search results can be displayed by housing complex unit or by apartment building.

The color of the marker shown on the screen differs depending on the search scheme. **Fig. 16** shows the marker color information. In the case of a search by value, the color is determined by the test result value from the learning model. In the case of a search by difference, the color is determined by the size of the difference between the test results based on the learning model and the actual value.

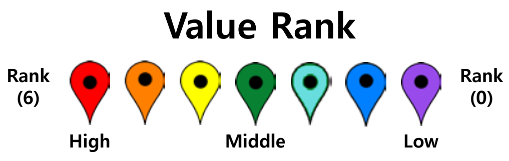


Fig. 16. Marcker Color Information

When a user searches by value, the results of the search appear based on the predicted rank, the test result obtained from the selected learning model. In exceptional cases, the actual value rank of the relevant region is used for the search result. This situation occurs when the actual value is assigned to the learning model. Fig. 17 depicts a screen showing a search by value, and the text below describes the components of the screen.

[1] Visualization map: shows the search result

[2] Marker information: shows the address of the relevant building and value rank when the marker is clicked

[3] Selected attribute information: shows the selected learning, test city, and unit search range

[4] Search city setting: selects the learning and test city

[5] Unit range setting: selects the search range of the unit (min, max)

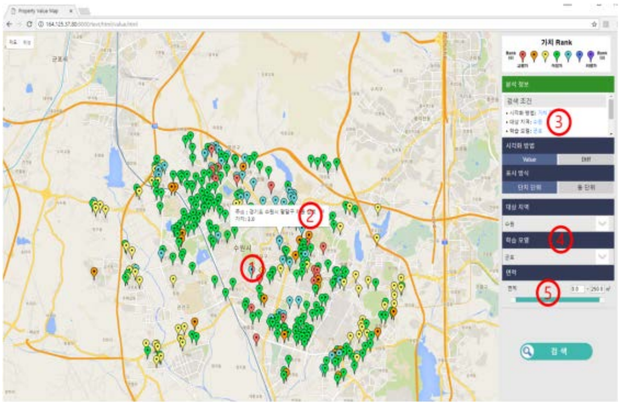


Fig. 17. Search by value attribute screen

The search by difference function displays the search results based on the difference between the actual value rank of the relevant region and the predicted rank that comesresulting from the test result of the relevant region using the selected learning model. Fig. 17 depicts the screen that showings the search by difference function, and the text. bBelow describes the components of the screen.

[1] Visualization map: shows prints the search result

[2] Marker information: prints the address of the relevant building and value score when the marker is clicked

[3] Selected attribute information: prints the selected attribute and unit search range

[4] Search condition: selects value attributes and the range of the unit (min, max)

[5] Weight condition: assigns search weight by attributes

The visualization service forof the building value attributes includes search by price per m^2 and search by value attributes (physical attributes, accessibility, and surrounding environment). Each search scheme is composed such that the search results can be expressed in complex units and apartment building units.

The user interface for of the web screen is composed of a default screen, search by price per m^2, and search by value attributes. Fig. 18 shows the screen forof searching by value attributes.

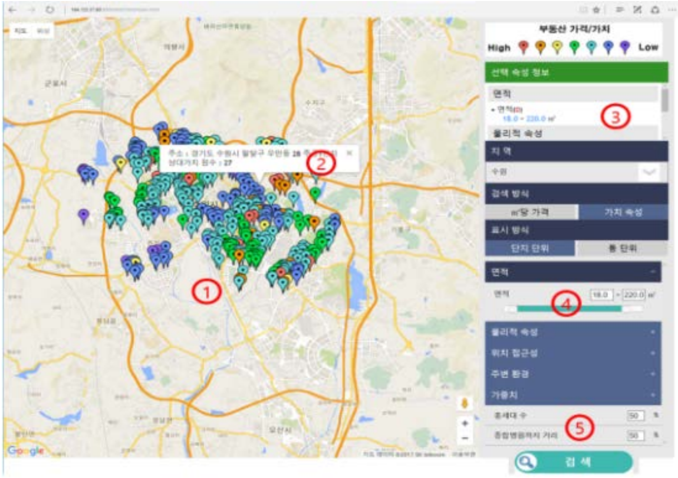


Fig. 18. Search by difference attribute screen

Therefore, the machine learning based prediction UI/UX will be very useful in terms of the building consumers. If the input data and UI/UX are more improvement, this system gives more information to consumers before buying buildings.

8. Conclusions

National organizations, local governments, and public institutions have recently released accumulated data from a variety of fields, including education, health and medical services, public administration, social welfare, and science and technology. This opening has catalyzed the creation of new businesses and jobs that make use of such data. In this paper, data regarding building value attributes were obtained by integrating recently released public data. This study implemented an objective and quantitative analysis regarding the value of real estate buildings by applying the RF and DNN scheme to the fabricated data set. In the experimental results, the RF scheme showed better performance than the DNN scheme. However, both the RF and the DNN scheme showed performance below 50% when the model of one region was applied to the other region.

References

[1] Onnara Real Estate, Article (CrossRef Link)
[2] Real Estate Transaction Management System, Article (CrossRef Link)
[3] Korea Appraisal Board, Article (CrossRef Link)
[4] Real Estate 114, Article (CrossRef Link)
[5] Wise Net, Article (CrossRef Link)
[6] Naver Real Estate, Article (CrossRef Link)
[8] LOBIG, Article (CrossRef Link)
[9] HUD, Article (CrossRef Link)
[10] UK Government services and information, Article (CrossRef Link)
[11] Japan Real Estate Institute, Article (CrossRef Link)

[12] Sang-koo Kang, “Public Information DB Open Space Platform for Information Integration and Integration,” *Real Estate Focus*, May 2016. Article (CrossRef Link).
[13] Jae-heon Sim, "A Study on the Estimation of Real Estate Value Using Machine Learning Algorithm," *Real Estate Focus*, Sep. 2016. Article (CrossRef Link).
[14] Young-im Cho, "Big data-based artificial intelligence using real estate and its prospects," *Real Estate Focus*, Sep. 2016. Article (CrossRef Link).
[15] Won-Gu Jung, "A Study on Prediction of Apartment House Price Index Using Artificial Neural Network," *Housing Research*, Mar. 2007. Article (CrossRef Link).
[16] Gyu-pil Yeon, "A Machine Learning Approach to Improve the Appropriateness of Standardized Housing Price," *Real Estate Research*, June 2015. Article (CrossRef Link).
[17] Chang-ro Lee and Ki-ho Park, "Application of Machine Learning Model for Estimation of Apartment Price," *Journal of Korean Geographical Society*, Apr. 2016. Article (CrossRef Link).
[18] Hwang-jong Seong, "Artificial intelligence era, realization of complex combination of real estate big data," *Real estate focus*, Apr. 2017. Article (CrossRef Link).
[19] Selenium HQ, Article (CrossRef Link)
[20] PhantomJS, Article (CrossRef Link)
[21] Arch GIS, Article (CrossRef Link)
[22] EDS Viewer, Article (CrossRef Link)

Woosik Lee received the B.S. degree in Computer Science from the Kyonggi University, Korea, in 2009, and the M.S. degree and the Ph.D. degree in the Computer Science from Kyonggi University in 2011 and 2016. In 2016, he was a research member of Korea Institute of Civil engineering and building Technology. In 2017, he was an assistant professor at the department of computer science, Kyonggi University, Korea. Since 2018, he has been a vice researcher at Social Security Information Service. His research interests include wireless systems, sensor networks, internet of things, and energy management protocols.

Namgi Kim received the B.S. degree in Computer Science from Sogang University, Korea, in 1997, and the M.S. degree and the Ph.D. degree in Computer Science from KAIST in 2000 and 2005, respectively. From 2005 to 2007, he was a research member of the Samsung Electronics. Since 2007, he has been a faculty of the Kyonggi University. His research interests include sensor system, wireless system, cloud computing, SDN, and mobile platform.

Yoon-Ho Choi is a faculty member at the School of Computer Science & Engineering in Pusan National University, Busan, Korea. He received his M.S. and Ph.D. degrees from the School of Electrical and Computer Engineering, Seoul National University, S. Korea, in Aug. 2004 and Aug. 2008, respectively. He was a postdoctoral scholar at Seoul National University, Seoul, S. Korea from Sep. 2008 to Dec. 2008 and in Pennsylvania State University, University Park, PA, USA from Jan. 2009 to Dec. 2009. He worked as a senior engineer at Samsung Electronics from May 2010 to Feb. 2012. He was a faculty member at the Department of Convergence Security in Kyonggi University from Mar. 2012 to Aug. 2014. He has served as a TPC member in various international conferences and journals. His research interests include Deep Packet Inspection (DPI) for high-speed intrusion prevention, mobile computing security, vehicular network security for realizing secure computer and networks.

Yong Soo Kim is Assistant Professor at the Department of Industrial and Management Engineering, Kyonggi University, Korea. He received B.S., M.S. and Ph.D. degree in industrial engineering from KAIST, respectively. His research interests include decision support system, data mining, and social network analysis. Nowadays, his research topics are focused on recommender system for mobile internet.

Byoung-Dai Lee is an assistant professor at the department of computer science, Kyonggi University, Korea. He received his B.S. and M.S. degrees in Computer Science from Yonsei University, Korea in 1996 and 1998 respectively. He received his Ph.D. degree in Computer Science and Engineering from University of Minnesota, Minneapolis, U.S.A. in 2003. Before joining the Kyonggi University, he worked at Samsung Electronics, Co., Ltd as a senior engineer from 2003 to 2010. His research interests include cloud computing, mo