Proposing Machine Learning Models Suitable for Predicting Open Data Utilization

Junyoung Jeong and Keuntae Cho *

Posted Date: 19 June 2024

doi: 10.20944/preprints202406.1319.v1

Keywords: open data; open government data; open data utilization

 doi:10.20944/preprints202406.1319.v1

Article

Proposing Machine Learning Models Suitable for Predicting Open Data Utilization

Junyoung Jeong [1] and Keuntae Cho [2,*]

[1] Graduate School of Management of Technology, Sungkyunkwan University, 2066 Seobu-Ro, Jangan-Gu, Suwon 16419, Republic of Korea; j.y.jeong@skku.edu
[2] Department of Systems Management Engineering & Graduate School of Management of Technology, Sungkyunkwan University, 2066 Seobu-Ro, Jangan-Gu, Suwon 16419, Republic of Korea
* Correspondence: ktcho@skku.edu; Tel.: +82-031-290-7602

Abstract: As the digital transformation accelerates in our society, open data is being increasingly recognized as a key resource for digital innovation in the public sector. This study explores the following two research questions: 1) Can a machine learning approach be appropriately used for measuring and evaluating open data utilization? 2) Should different machine learning models be applied for measuring open data utilization depending on open data attributes (field and usage type)? This study used single-model (Random Forest, XGBoost, LightGBM, CatBoost) and multi-model (Stacking Ensemble) machine learning methods. A key finding is that the best-performing models differed depending on open data attributes (field and type of use). The applicability of the machine learning approach for measuring and evaluating open data utilization in advance was also confirmed. This study contributes to open data utilization and to the application of its intrinsic value to society.

Keywords: open data; open government data; open data utilization

1. Introduction

Since the onset of the COVID-19 pandemic, digital transformation has accelerated in society and open data has garnered attention as a key resource in digital conversion within the public sector [1]. Concurring with this trend, discussions regarding the development of information and communications technology have accelerated [2], and the potential of open data, as a digital asset and resource, to contribute to social and economic growth through sustainable value creation has been mentioned [3]. In the context of sustainable environmental, social, and governance management, Helbig et al. [4] noted that open data use can provide significant strategic value to organizations and increase business efficiency.

In response, governments—as key data providers—worldwide have pursued policies for open government data. Since the initiation in the United Kingdom and the United States of America of the development of open government policies aimed at open data openness and transparency enhancements in 2009, many countries, including South Korea, have endeavored to utilize data-based innovations to increase transparency [5–7]. According to evaluations by the Organization for Economic Co-operation and Development on the Open-Useful-Reusable DATA index, South Korea achieved the top position for four years (i.e., 2015, 2017, 2019, and 2023), scoring 0.91 points (out of 1) in 2023. As of March 2024, the South Korean government has facilitated the opening of over 87,000 diverse datasets from 1,031 organizations through the Open Data Portal (data.go.kr).

However, there are limitations to integrating the quantitative outcomes of open government data with qualitative utilization outcomes [8]. The intrinsic value of open data lies not merely in the openness of the data itself, but in the added value that can be derived from its application to businesses. This means that open data can generate substantial value when utilized by end users [9], making efforts to assess and improve open data value and to render users aware of open data utility critical endeavors [10–12].

 doi:10.20944/preprints202406.1319.v1

Various studies have been conducted on open data utilization, but most have focused on indirect and complementary factors (e.g., legal frameworks, governance, policy, and technology) rather than on the intrinsic value of open data utilization. For example, there are various studies on open data legal frameworks and governance, which have emphasized the role and structure of relevant legislation in our rapidly changing society, highlighting the need for well-established legal frameworks [13–15]. In another study, trust in open data governance has been emphasized as a key factor of open data utilization [16]. In the context of South Korea's open data utilization policies, a study compared the open data policies of different administrations, finding an initial focus on the openness and scalability of open data utilization in diverse areas [17]. Research on open data technology have also provided technical recommendations for the enhancement of platform infrastructure, along with insights for analytical tools and technical standards [10,18–22]. Thus, previous studies have primarily offered results relevant for post-prescriptive alternatives regarding open data utilization, whereas systematic approaches to measure and evaluate the intangible value of open data utilization have not yet been established, and our understanding of how to proactively tackle the matter is limited.

2. Related Research

2.1. Research on Open Data Utilization

Schumpeter et al. [23] argued that innovations such as new products, production methods, markets, and economic organizations are both directly and indirectly related to open data utilization [24,25]. Meanwhile, open data has been recognized as a resource for innovation that can create additional value based on its intrinsic characteristics [26,27]. Various studies have confirmed that active open data use is beneficial for researchers, companies, and other stakeholders, including for creating new businesses and supporting alternative decision-making [28,29]. There are also many explorations of the potential of open data utilization to foster innovation in various sectors of private enterprise products and services [25,30]. Regarding the economic impact of open data, the estimates point to tens of billions of euros annually [31].

According to the South Korean government, among the companies that utilized open data and were surveyed (n = 1,003), a considerable number (63.7%) reported that open data plays a vital role in their business [32]. Janssen et al. [33] further demonstrated that open data utilization can activate data-business linkages, support decision-making, and enhance business quality. This points to the indispensability of open data utilization as a trend for leading nations and companies in the era of artificial intelligence. In addition, research on legal frameworks and governance regarding open data utilization note that while open data should contribute to a range of social and political goals, the release of government data containing personal information can threaten people's privacy and related rights and interests; accordingly, there have been proposals for frameworks that consider privacy risks in open data utilization in the public sector [34]. Thompson et al. [35] suggested organizational governance adjustments tailored to unique situations, and emphasized the need for strong partnerships between information technology professionals and data specialists when actively using open data.

In a policy study related to open data utilization, Zuiderwijk et al. [36] noted that governments aim to promote and induce data release and gain benefits from data utilization when establishing open data policies. These cited authors also developed a framework for comparing individual policies based on the open data policies of Dutch government agencies. Meanwhile, Bertot et al. [37] noted the need for a robust data sharing and interoperability framework as big data and open data are being increasingly exchanged in real time.

Research on open data utilization-related technology is also underway. Regarding technological infrastructure (e.g., data platforms), Máchová et al. [38] evaluated national open data portal usability based on data explorability, accessibility, and reusability. Osagi et al. [39] noted the extensibility of open data platforms and suggested improvements to enable citizens and civil society organizations to effectively utilize open data. Furthermore, research has been conducted on other technical aspects

 doi:10.20944/preprints202406.1319.v1

related to open data utilization, such as data quality, with Vetrò et al. [40] pointing out that opening data without proper quality management could diminish data reuse and influence negative usability.

In summary, the literature presents trends toward the advancement of open data utilization from diverse perspectives. The major limitation of the related studies is that they only deliver indirect alternatives for measuring and evaluating open data utilization value. Therefore, this study proposes a discriminative alternative that directly measures and evaluates the intangible value of open data utilization based on the characteristics of the open data managed by the South Korean government.

2.2. Research on Machine Learning Application

In this study, we focused on machine learning as a quantitative tool for evaluating open data utilization. Machine learning is already being applied in various fields related to everyday life, such as the medical sector [41–43], environmental sector [44–47], and construction industry [48]. The introduction of machine learning into these fields has greatly expanded the scope and number of insights as to the application of theoretical knowledge in the real world. Machine learning has also customarily been used as a tool to predict intangible values (e.g., through risk assessment) in the manufacturing industry [49,50].

Regarding the use of machine learning as a tool, several studies have used machine learning to, for instance, predict citation counts, as a direct indicator of patent utilization (i.e., high citation counts indicate highly-utilized patents). Indeed, patents in the top 1% of citation counts are defined as impactful breakthrough inventions relevant to commercialization and future technology development [51]. Researchers have also used machine learning to classify utility through classification algorithms such as SOM, KPCA, and SVM [52]; predict patent citation counts using boosting algorithms (e.g., XGBoost classifier) [53]; predict utilizable patents for research and development investment decisions in companies [54]. Therefore, machine learning tools can be considered appropriate to measure and evaluate the utilization value of intangible assets such as patents.

There has also been considerable research on training data for machine learning. Concerning the characteristics of training data, related studies have mentioned that tree-based algorithm utilization is appropriate when including a large amount of categorical data in the training dataset [55,56]. For example, a study related to construction waste generation suggested that the development of machine learning prediction models for small-scale datasets composed of categorical variables could be improved by applying random forest and GBM algorithms [55,56]. Corroborating these assertions, another research [57] mentioned the superiority of random forest over GBM in terms of stability and accuracy for small-scale datasets composed of categorical variables, demonstrating performance differences that reflect data characteristics.

Regarding training data size, it is generally known from previous research that large-scale (vs. small-scale) training datasets provide better performance. However, Ramezan et al. [58] demonstrated performance differences depending on the algorithm as training data size was adjusted, highlighting the importance of specific situational considerations. Based on prior research on training data, this study attempted to compare model performance by distinguishing between training based on an integrated dataset without field distinctions and on 16 field-specific datasets. The goal was confirming the potential performance differences according to training data size across the 16 fields.

There is a plethora of research on machine learning algorithm performance. In a comparison of the SVM, random forest, and ELM algorithms when predicting intrusion detection showcased the superiority of the ELM model in terms of performance [59]. In a research related to early diabetes diagnosis [60], after analyzing 66,325 patient records based on 18 risk factors, logistic regression was deemed the superior model. Moreover, when predicting the bending strength of steel structures and the bonding strength of surfaces [61], the ANN algorithm seemed to have superior performance based on a comparison with other algorithms (i.e., random forest, XGBoost, and LightGBM). A research [62] predicting the residual value of construction machinery compared the coefficients of

 doi:10.20944/preprints202406.1319.v1

determination for LightGBM, XGBoost, and MDT, reporting that the MDT algorithm was the best prediction model (accuracy of 0.9284).

Concerning machine learning stacking ensembles (i.e., a method of combining individual prediction results to enhance final performance), a study predicting corporate bankruptcy used basic models (e.g., KNN, Decision Tree, SVM, and Random Forest) based on the financial data of companies to show that the stacking ensemble model with LightGBM achieved an accuracy of over 97% [63]. In a study predicting harmful algal blooms (HABs), the base models of XGBoost, LightGBM, and CatBoost, and Linear Regression were used to construct a metamodel, which in turn confirmed the applicability of the stacking ensemble technique. Based on these previous studies, the current research conducted comparative analyses of single models and stacking ensemble techniques to enhance model performance.

3. Materials and Methods

3.1. Data

In this study, we delimited our scope to the structured data provided by the South Korean government from 2012–2022, which is openly available at Open Data Portal (data.go.kr). We collected and analyzed metadata through the download method (File Data) and the API method (OpenAPI Data), excluding unusable variables such as contact information (e.g., the responsible person's name and phone number). The metadata includes detailed information, including related fields and data descriptions. This study considered that the amount of data varied across the 16 fields depending on the scope of the South Korean government's administrative work; thus, the experiments were conducted while dividing the training dataset into an integrated dataset without field division, and 16 field datasets with field division. We evaluated the performance of the models separately for each case. The utilized data consisted of metadata for 44,648 File Data and 6,677 OpenAPI Data after removing duplicates and missing values.

3.1.1. Input Variables

This study separated the training datasets of File Data and OpenAPI Data depending on the utilization method. The File Data metadata comprised 37 variables, among which 23 were continuous and 14 were categorical variables. The OpenAPI Data metadata is accessible in real-time via API calls, and comprised 42 variables, among which 23 were continuous and 19 were categorical variables.

3.1.2. Target Variables

To establish the target variables that quantitatively represent open data utilization [64], we constructed indicators with a normal distribution, which served to consider model performance in the analyses [65–67]. We then verified the normal distribution of each target variable. For File Data, we adjusted the number of downloads based on the provision period and considered the number of attachments, as shown in Figure 1(a). The provision period incorporates the concept of patent citation half-life, reflecting an adjustment for the period of utilization in the patent field. For OpenAPI Data, we utilized the number of API calls and API utilization requests as the target variables. Similar to procedures for the File Data dataset, we established target variables considering the service period and confirmed whether they followed a normal distribution, as illustrated in Figure 1(b) [67].

$$\text{(a) } Target\ variable\ of\ File\ Data = \log\left(\frac{Download\ counts}{provision\ period \times Number\ of\ attachments}\right)$$

$$(b)\ Target\ variable\ of\ OpenAPI\ Data = \log\left(\frac{API\ call\ counts \times API\ request\ counts}{service\ period}\right)$$

Figure 1. Target variable of File Data (a) and OpenAPI Data (b).

3.2. Proposed Methods

 doi:10.20944/preprints202406.1319.v1

3.2.1. Single Model Methods

As the training dataset contained many categorical variables, we selected tree-based algorithms (Random Forest, XGBoost, LightGBM, and CatBoost) for use in this study [68–70].

- Random Forest: Random forest algorithms are primarily composed of decision trees, and the results of these trees are summed to produce a final result that maximizes algorithm performance [71]. The advantage of decision tree analysis is the easy and intuitive understanding it provides as the results are presented in a single tree structure; the disadvantage is its lower predictiveness owing to the consideration of only one predictor when dividing the tree branches. Small data changes can also transform tree composition [71,72]. Therefore, while decision tree analysis has a relatively low bias, it has a high variance error, rendering model generalization more difficult. A machine learning algorithm used to overcome this weakness is random forest, which analyzes and aggregates multiple decision trees to form a forest of randomly sampled decision trees, which then is used to create a final prediction model. Random forest algorithms iteratively create independent decision trees to maximize sample and variable selection randomness, hence reducing prediction error by lowering variance and sustaining a low bias in the decision tree [71,72]. When using data with multiple explanatory variables, random forest algorithms also provide stability by considering interactions and nonlinearities between the explanatory variables. A visual representation of a random forest model is presented in Figure 2(a).
- eXtreme Gradient Boosting (XGBoost): XGBoost is an algorithm proposed by Chen et al. [73] for use with large-scale datasets, and is meant to compensate for overfitting issues while improving stability and training speed. XGBoost is known for its performance and effectiveness owing to the implementation of the gradient boost learning technique, which is a well-known technique in machine learning. Specifically, it uses a greedy algorithm to construct the most optimal model and improve weak classifiers, and this occurs while controlling complexity using distributed processing to compute optimal weights; this all serves to minimize learning loss and overfitting. This algorithm can be trained on categorical and continuous data, and each leaf contributes to the final score of the model; its analysis procedure is as follows: 1) measure the accuracy of the generated tree classifiers; 2) randomly generate strong-to-weak classifiers in each order; 3) sequentially improve the classifiers to generate a strong tree classifier. XGBoost proceeds to the max_depth parameterized during training, and then prunes in reverse if the improvement in the loss function does not reach a certain level [73]. During this process, the model can be pruned to remove unnecessary parts of the tree classifier and prevent overfitting. A visual representation of the XGBoost algorithm is presented in Figure 2(b).
- Light Gradient Boosting Machine (LightGBM): developed by Microsoft, this model uses the leaf-wise partitioning method to create highly-accurate models [74]. It is based on the gradient boosting decision tree ensemble learning technique, which has the advantage of dividing the branches at each node based on the best-fit nodes. It uses this learning technique with various algorithms to reduce the number of dimensions of individual data. This technique uses level wise for horizontal growth and the traverse of the nodes of the decision tree preferentially from the root node. For vertical growth, it splits at the node with the largest maximum delta loss, assuming that the loss can be further reduced by growing the same leaf. Furthermore, the two methods used in LightGBM to reduce the number of samples and features are gradient-based one-side sampling and exclusive feature bundling. Gradient-based one-side sampling is an under-sampling technique guided by the training set's skewness, considering that samples with a larger skewness in absolute value contribute more to learning; accordingly, those with a smaller gradient are randomly removed. A visual representation of the LightGBM algorithm is shown in Figure 2(c).
- Categorical Boosting (CatBoost): This is is a library based on gradient boosting [75]. CatBoost performs well with categorical data [76], and processes them using the statistics of the target values while converting each categorical variable to a number. Thus, it is a great performer for most machine learning tasks that require categorical data processing [77]. In a past study, CatBoost performed better than other gradient boosting libraries because of its ability to handle categorical variables and its optimized algorithm [76]. As aforementioned, it converts categorical

 doi:10.20944/preprints202406.1319.v1

variables into numbers using various methods, implying the non-need for the preprocessing of categorical data and the possibility of directly processing it using this algorithm [75]. CatBoost also uses multiple strategies to avoid overfitting, which is a common problem in gradient boosting [77]. This algorithm is hence primarily used as a tool for solving classification and regression problems, performing particularly well on categorical data-related problems [78]. A visual representation of the CatBoost algorithm is presented in Figure 2(d).

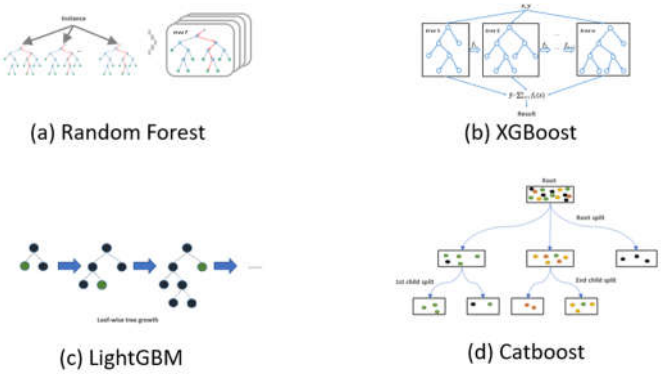

Figure 2. Concept of the single model algorithms used in this study.

3.2.2. Multi-Model Method (Stacking Ensemble)

In this study, we utilized the multi-model stacking ensemble method in addition to the single model method. Stacking ensembles are constructed by combining various different single models to achieve better performance, enabling the use of the strengths of each algorithm and the compensation for their corresponding weaknesses. That is, it aims to create a better performing model by combining different models [79]. In this study, single models (random forest, XGBoost, LightGBM, and CatBoost) were used to form a stacking ensemble, and GBM was used as the metamodel for the final prediction [67,80,81]. Figure 3 shows a diagram of the stacking ensemble and multi-model method utilized in this study.

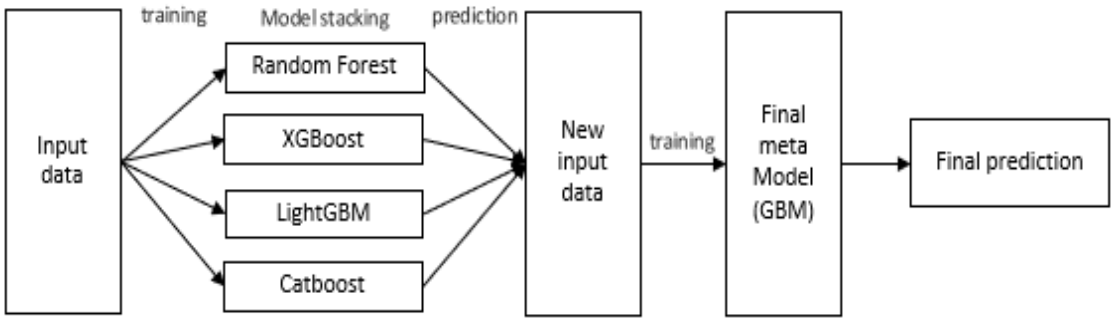

Figure 3. Conceptualization of the multi-model method (stacking ensemble) used in this study.

3.3. Model Performance Evaluation

To evaluate model performance, we utilized the mean squared error (MSE) and root mean squared error (RMSE) metrics to measure error size (i.e., the difference between the predicted and actual values). Because MSE squares the difference between the predicted and actual values, it is sensitive to outliers, meaning that if the predicted value differs from the actual value by a large amount, the difference that is yielded will be relatively large. We hence decided to also implement RMSE to compare the tendency of the effect on outliers. The formula used is shown in Figure 4(a).

 doi:10.20944/preprints202406.1319.v1

$$\text{(a) } MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \ (y_i = predicted\ value, \hat{y}_i = \ actual\ value)$$

$$\text{(b) } RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \ (y_i = predicted\ value, \hat{y}_i = \ actual\ value)$$

Figure 4. Formula of mean squared error (a) and root mean squared error (b).

RMSE is an indicator of scale-dependent errors, and is organized as shown in Figure 4(b). It tends to increase when the magnitude of the value to be predicted is large, and decrease when the magnitude of the value to be predicted is small. This metric is used to evaluate the prediction performance of a model by calculating the square root of the squared error [82], and in so doing, it has the same units as the error value [83]—unlike MSE. Thus, the metric can give researchers an intuitive idea of the average size of the error between the predicted and true values [82]. Both MSE and RMSE are widely used in regression models in machine learning to evaluate model predictive performance [83]. Specifically, the smaller the MSE and RMSE, the better the model's prediction performance. We conducted a comparative analysis of MSE and RMSE between the single model and multi-model methods. The workflow for model performance evaluation is illustrated in Figure 5.

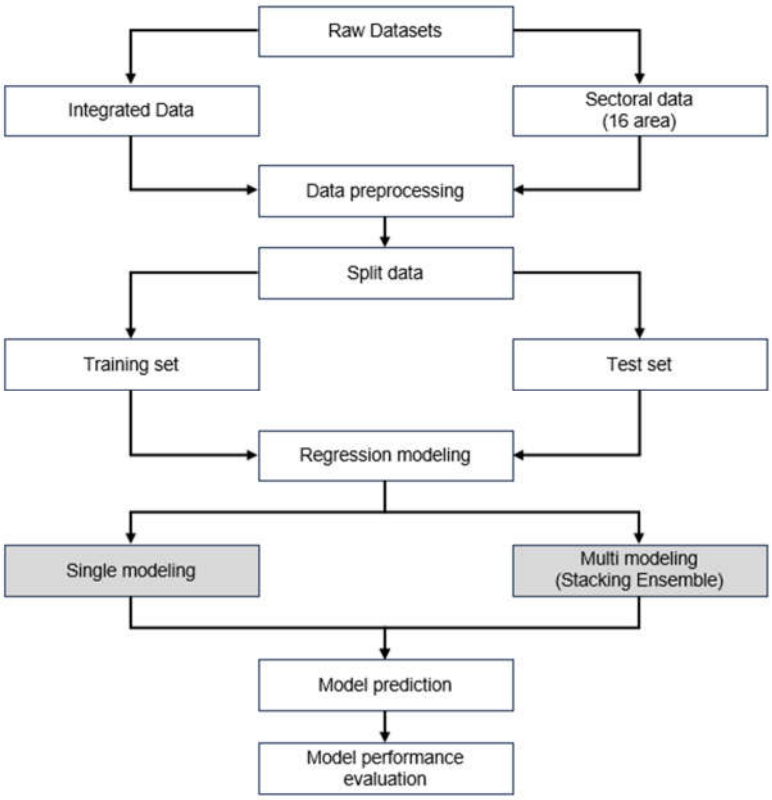

Figure 5. Workflow for predicting open data utilization using machine learning models.

4. Results

4.1. File Data

In the File Data, similar trends were observed for MSE and RMSE (Tables 1, 2, and 3). The range of MSE was 0.256–0.793 for the single model and 0.314–0.763 for the multi-model method; for RMSE, the ranges were 0.506–0.891 and 0.560–0.873, respectively. Regarding the single model method, the following ranges were identified for the algorithms: for random forest, MSE from 0.530–0.793 and

 doi:10.20944/preprints202406.1319.v1

RMSE from 0.592–0.891; for XGBoost, 0.313–0.754 and 0.560–0.868, respectively; for LightGBM, 0.313–0.736 and 0.559–0.858, respectively; for CatBoost, 0.256–0.678 and 0.506–0.824, respectively.

Table 1. Superior models, training data, and performance metrics by field using File Data.

Item	Model	Training Data	MSE	RMSE
Public Administration	Catboost	Integrated	0.665	0.443
Science & Technology	Catboost	Field	0.718	0.516
Education	Catboost	Integrated	0.625	0.390
Transportation & Logistics	Catboost	Integrated	0.643	0.414
Land Management	Catboost	Integrated	0.598	0.358
Agriculture & Fisheries	Catboost	Integrated	0.652	0.425
Culture & Tourism	Stacking Ensemble	Integrated	0.701	0.491
Law	Catboost	Integrated	0.506	0.256
Healthcare	Stacking Ensemble	Integrated	0.627	0.393
Social Welfare	Catboost	Integrated	0.638	0.407
Industry & Employment	Catboost	Integrated	0.637	0.405
Food & Health	Catboost	Integrated	0.584	0.341
Disaster Safety	Stacking Ensemble	Integrated	0.622	0.387
Finance	Catboost	Field	0.557	0.311
Unification & Diplomacy	Catboost	Integrated	0.623	0.388
Environment & Meteorology	Catboost	Integrated	0.671	0.450

Table 2. Comparison of model performance based on mean squared error by field using File Data.

Item		**Method**									
		Single Model								**Multi-Model (Stacking Ensemble)**	
	Algorithm	**Random Forest**		**XGBoost**		**LightGBM**		**CatBoost**			
	Training Data	**Integrated**	**Field**	**Integrated**	**Field**	**Integrated**	**Field**	**Integrated**	**Field**	**Integrated**	**Field**
Field	Public Administration	0.556	0.541	0.533	0.527	0.534	0.504	0.443	0.501	0.505	0.494
	Science & Technology	0.793	0.592	0.754	0.555	0.736	0.616	0.678	0.516	0.763	0.607
	Education	0.478	0.528	0.472	0.559	0.470	0.503	0.390	0.484	0.450	0.509
	Transportation & Logistics	0.538	0.631	0.552	0.621	0.553	0.657	0.414	0.593	0.511	0.618
	Land Management	0.534	0.501	0.557	0.495	0.549	0.490	0.358	0.488	0.516	0.523

 doi:10.20944/preprints202406.1319.v1

	Agriculture & Fisheries	0.482	0.491	0.478	0.468	0.478	0.445	0.425	0.443	0.435	0.470
	Culture & Tourism	0.535	0.581	0.519	0.539	0.546	0.552	0.501	0.557	0.491	0.544
	Law	0.385	0.374	0.335	0.344	0.313	0.408	0.256	0.326	0.347	0.373
	Healthcare	0.435	0.523	0.429	0.496	0.416	0.508	0.466	0.504	0.393	0.509
	Social Welfare	0.486	0.483	0.433	0.487	0.450	0.479	0.407	0.480	0.433	0.461
	Industry & Employment	0.497	0.515	0.476	0.519	0.463	0.471	0.405	0.478	0.453	0.476
	Food & Health	0.544	0.518	0.515	0.429	0.518	0.450	0.341	0.435	0.500	0.475
	Disaster Safety	0.451	0.663	0.416	0.698	0.424	0.678	0.525	0.631	0.387	0.635
	Finance	0.360	0.350	0.348	0.313	0.356	0.329	0.396	0.311	0.330	0.314
	Unification & Diplomacy	0.475	0.465	0.513	0.528	0.525	0.556	0.388	0.498	0.475	0.532
	Environment & Meteorology	0.528	0.601	0.519	0.622	0.517	0.591	0.450	0.575	0.482	0.577

Table 3. Comparison of model performance based on root mean squared error by field using File Data

Item		**Method**									
		Single Model								**Multi-Model (Stacking Ensemble)**	
Algorithm		**Random Forest**		**XGBoost**		**LightGBM**		**CatBoost**			
Training Data		**Integrated**	**Field**	**Integrated**	**Field**	**Integrated**	**Field**	**Integrated**	**Field**	**Integrated**	**Field**
Field	Public Administration	0.745	0.735	0.730	0.726	0.731	0.710	0.665	0.708	0.710	0.703
	Science & Technology	0.891	0.769	0.868	0.745	0.858	0.785	0.824	0.718	0.873	0.779
	Education	0.692	0.727	0.687	0.748	0.686	0.710	0.625	0.696	0.671	0.714

 doi:10.20944/preprints202406.1319.v1

Transportation & Logistics	0.733	0.794	0.743	0.788	0.743	0.811	0.643	0.770	0.715	0.786
Land Management	0.731	0.708	0.746	0.704	0.741	0.700	0.598	0.698	0.718	0.723
Agriculture & Fisheries	0.694	0.700	0.691	0.684	0.691	0.667	0.652	0.666	0.660	0.686
Culture & Tourism	0.732	0.762	0.721	0.734	0.739	0.743	0.708	0.746	0.701	0.738
Law	0.620	0.611	0.579	0.587	0.559	0.639	0.506	0.571	0.589	0.611
Healthcare	0.660	0.723	0.655	0.705	0.645	0.713	0.683	0.710	0.627	0.713
Social Welfare	0.697	0.695	0.658	0.698	0.671	0.692	0.638	0.693	0.658	0.679
Industry & Employment	0.705	0.717	0.690	0.720	0.680	0.686	0.637	0.692	0.673	0.690
Food & Health	0.738	0.720	0.718	0.655	0.720	0.671	0.584	0.660	0.707	0.689
Disaster Safety	0.672	0.814	0.645	0.835	0.651	0.823	0.725	0.794	0.622	0.797
Finance	0.600	0.592	0.590	0.560	0.597	0.573	0.629	0.557	0.575	0.560
Unification & Diplomacy	0.689	0.682	0.717	0.727	0.724	0.746	0.623	0.705	0.689	0.730
Environment & Meteorology	0.727	0.776	0.721	0.788	0.719	0.769	0.671	0.758	0.694	0.760

Regarding the performance of the single model and the multi-model methods, the single model method excelled in 13 fields, whereas the multi-model method showed superior performance in three fields. For the single model method, CatBoost exhibited superior performance across all 13 fields. When the single model method included CatBoost trained on the integrated data, the model was the best performing in 11 fields (i.e., Public Administration, Education, Transportation & Logistics, Land Management, Agriculture & Fisheries, Law, Social Welfare, Industry & Employment, Food & Health, Unification & Diplomacy, Environment & Meteorology); when trained on field-specific data, it showed the best performance in the Science & Technology and Finance fields. For the multi-model method, it showed superior performance in three fields (Culture & Tourism, Healthcare, Disaster & Safety) when trained on integrated data.

Comparing the performance of each field based on MSE and RMSE, the field with the best performance was Law (MSE, 2.533; RMSE, 1.592), while the field with the poorest performance was Transportation & Logistics (MSE, 14.204; RMSE, 3.769). In the Law Field, the best performance was observed with the multi-model method trained on integrated data. In the Transportation & Logistics field, the best performance was observed with the single model method with LightGBM trained on integrated data.

When comparing the performance of models within the same field, the Law field showed the largest performance difference (among the 16 fields) across models. Specifically, based on RMSE, the

 doi:10.20944/preprints202406.1319.v1

performance of the best model was approximately 4.378 times superior to that of the poorest model; based on MSE, the performance of the best model was approximately 19.166 times superior. Meanwhile, the Public Administration field showed the smallest performance difference across models. When considering RMSE, the performance of the best model was approximately 1.089 times superior to that of the poorest model; according to MSE, it was approximately 1.185 times superior. A summary of the superior algorithms, training data, and performance metrics (MSE and RMSE) for each field is presented in Table 1.

4.2. OpenAPI Data

For OpenAPI Data, similar trends were observed for the MSE and RMSE metrics (Tables 4, 5, and 6), as occurred for the File Data. The MSE ranges were 2.906–48.547 for the single model and 2.533–35.837 for the multi-model method, whereas the RMSE ranges were 1.705–6.968 and 1.592–5.986, respectively. Regarding the single model method, the following ranges were identified for the algorithms: for random forest, MSE from 2.906–29.163 and RMSE from 1.705–5.400; for XGBoost, 4.239–48.547 and 2.059–6.968, respectively; for LightGBM, 4.259–26.606 and 2.064–5.158, respectively; for CatBoost, 4.332–28.268 and 2.081–5.317, respectively. The difference between the results from OpenAPI Data and File Data was that there was a wider range of superior model algorithms in the OpenAPI Data. Regarding the performance of the single model and multi-model methods, the single model method excelled in 13 fields, while the multi-model method showed superior performance in three fields. For the single model method, the best performance was achieved in six fields (Public Administration, Science & Technology, Culture & Tourism, Disaster Safety, Finance, Unification & Diplomacy) with the random forest algorithm trained on field-specific data.

Table 4. Superior models, training data, and performance metrics by field using OpenAPI Data.

Item	Model	Training Data	MSE	RMSE
Public Administration	Random Forest	Field	3.130	9.797
Science & Technology	Random Forest	Field	3.082	9.499
Education	XGBoost	Integrated	2.206	4.867
Transportation & Logistics	LightGBM	Integrated	3.769	14.204
Land Management	LightGBM	Field	2.986	8.915
Agriculture & Fisheries	Stacking Ensemble	Integrated	2.645	6.995
Culture & Tourism	Random Forest	Integrated	2.326	5.412
Law	Stacking Ensemble	Integrated	1.592	2.533
Healthcare	Stacking Ensemble	Integrated	2.048	4.193
Social Welfare	Catboost	Field	2.290	5.244
Industry & Employment	XGBoost	Integrated	2.725	7.427
Food & Health	LightGBM	Field	2.622	6.875
Disaster Safety	Random Forest	Field	2.140	4.580
Finance	Random Forest	Field	3.453	11.921
Unification & Diplomacy	Random Forest	Field	2.463	6.069
Environment & Meteorology	Catboost	Integrated	2.093	4.380

Table 5. Comparison of model performance based on mean squared error by field using OpenAPI Data.

 doi:10.20944/preprints202406.1319.v1

Item		Method								
		Single Model							Multi-Model (Stacking Ensemble)	
Algorithm		Random Forest		XGBoost		LightGBM		CatBoost		
Training Data		Integrated	Field	Integrated	Field	Integrated	Field	Integrated	Field	Integrated	Field
Field	Public Administration	11.357	9.797	10.784	11.223	11.103	11.608	11.385	10.832	10.204	10.869
	Science & Technology	25.342	9.499	30.019	15.414	26.606	22.801	27.600	11.134	27.830	22.103
	Education	6.712	5.316	4.867	7.509	4.930	8.096	5.530	6.386	5.002	7.271
	Transportation & Logistics	17.987	18.608	16.565	18.705	14.204	22.423	17.206	16.982	15.825	19.322
	Land Management	11.695	10.516	13.604	10.062	11.494	8.915	11.877	10.045	13.185	10.850
	Agriculture & Fisheries	7.624	9.431	7.856	8.825	7.232	10.403	8.266	9.274	6.995	9.945
	Culture & Tourism	5.412	7.830	5.713	6.900	5.760	8.535	5.859	7.249	5.470	8.095
	Law	2.906	29.163	9.217	48.547	9.736	14.011	4.332	28.268	2.533	35.837
	Healthcare	5.347	6.348	4.239	7.688	4.259	6.668	4.972	6.672	4.193	4.843
	Social Welfare	6.825	5.335	6.534	6.437	7.095	7.565	6.845	5.244	6.295	6.590
	Industry & Employment	8.810	11.382	7.427	14.137	8.058	12.263	8.677	10.555	8.037	11.853
	Food & Health	8.227	7.005	7.594	7.805	8.309	6.875	9.990	7.347	7.677	12.030
	Disaster Safety	7.477	4.580	9.845	6.434	8.911	6.150	9.368	5.572	8.386	5.418
	Finance	18.654	11.921	20.085	16.463	17.981	14.147	17.611	14.247	19.444	14.862
	Unification & Diplomacy	12.790	6.069	8.613	9.562	8.792	7.075	7.125	7.535	6.480	7.565
	Environment & Meteorology	6.628	11.139	5.459	11.387	4.447	10.458	4.380	9.506	5.252	10.120

 doi:10.20944/preprints202406.1319.v1

Table 6. Comparison of model performance based on root mean squared error by field using OpenAPI Data.

Item		Method									
		Single Model								Multi-Model (Stacking Ensemble)	
Algorithm		Random Forest		XGBoost		LightGBM		CatBoost			
Training Data		Integrated	Field	Integrated	Field	Integrated	Field	Integrated	Field	Integrated	Field
Field	Public Administration	3.370	3.130	3.284	3.350	3.332	3.407	3.374	3.291	3.194	3.297
	Science & Technology	5.034	3.082	5.479	3.926	5.158	4.775	5.254	3.337	5.275	4.701
	Education	2.591	2.306	2.206	2.740	2.220	2.845	2.352	2.527	2.237	2.697
	Transportation & Logistics	4.241	4.314	4.070	4.325	3.769	4.735	4.148	4.121	3.978	4.396
	Land Management	3.420	3.243	3.688	3.172	3.390	2.986	3.446	3.169	3.631	3.294
	Agriculture & Fisheries	2.761	3.071	2.803	2.971	2.689	3.225	2.875	3.045	2.645	3.154
	Culture & Tourism	2.326	2.798	2.390	2.627	2.400	2.922	2.421	2.692	2.339	2.845
	Law	1.705	5.400	3.036	6.968	3.120	3.743	2.081	5.317	1.592	5.986
	Healthcare	2.312	2.519	2.059	2.773	2.064	2.582	2.230	2.583	2.048	2.201
	Social Welfare	2.613	2.310	2.556	2.537	2.664	2.750	2.616	2.290	2.509	2.567
	Industry & Employment	2.968	3.374	2.725	3.760	2.839	3.502	2.946	3.249	2.835	3.443
	Food & Health	2.868	2.647	2.756	2.794	2.882	2.622	3.161	2.711	2.771	3.468
	Disaster Safety	2.734	2.140	3.138	2.537	2.985	2.480	3.061	2.361	2.896	2.328
	Finance	4.319	3.453	4.482	4.057	4.240	3.761	4.197	3.774	4.410	3.855
	Unification & Diplomacy	3.576	2.463	2.935	3.092	2.965	2.660	2.669	2.745	2.546	2.751
	Environment & Meteorology	2.574	3.337	2.337	3.374	2.109	3.234	2.093	3.083	2.292	3.181

 doi:10.20944/preprints202406.1319.v1

For the Education and Industry & Employment fields, the best performance was achieved with the XGBoost model trained on integrated data. For the Transportation & Logistics field, the best performance was achieved with the LightGBM model trained on integrated data. For the Land Management and Food & Health fields, the best performance was achieved with the LightGBM model trained on field-specific data. In the Social Welfare field, the best performance was achieved with the CatBoost model trained on field-specific data. In the Environment & Meteorology field, the best performance was achieved with the CatBoost model trained on integrated data. In three fields (Agriculture & Fisheries, Law, Healthcare), the best performance was achieved using the multi-model method trained on integrated data.

Comparing the performance of each field based on MSE and RMSE, the field with the best performance was Law (MSE, 0.256; RMSE, 0.506), and that with the poorest performance was Science & Technology (MSE, 0.516; RMSE, 0.718). In the Law field, the best performance was achieved with the CatBoost algorithm trained on integrated data; in the Science & Technology field, the best performance was achieved when employing the CatBoost algorithm trained on field-specific data.

When comparing the performance of models within the same field, the Disaster & Safety field showed the largest performance difference (among the 16 fields) across models. Specifically, when considering RMSE, the performance of the best model was approximately 1.343 times superior to that of the poorest model; based on MSE, it was approximately 1.802 times superior. Meanwhile, the Agriculture & Fisheries field showed the smallest performance difference (among the 16 fields) across models. When considering RMSE, the performance of the best model was approximately 1.074 times superior to that of the poorest model; according to MSE, it was approximately 1.155 times superior. A summary of the superior algorithms, training data, and performance metrics (MSE and RMSE) for each field is presented in Table 4.

5. Discussion

Using open data metadata accumulated from 2012 to 2022 in South Korea, this study applied machine learning techniques to construct predictive models and propose an alternative approach for evaluating open data utilization in advance. Both single model (random forest, XGBoost, LightGBM, CatBoost) and multi-model (Stacking Ensemble) methods were applied. Considering the attributes of the open data used (fields and utilization methods), the training data were selectively utilized as integrated and field-specific data. The results showed that model and method (i.e., single model and multi-model methods) superiority varied by data attributes. This finding aligns with the research trends and outcomes mentioned by Si et al. [84], emphasizing the importance of considering data attributes in data analysis, as different attributes can influence model performance.

Regarding the implications of distinguishing between two open data utilization methods (i.e., File Data and OpenAPI Data), we observed that the distribution of the target variables was broader in the OpenAPI Data, and that the algorithms exhibiting superior performance were more diverse. This corresponds to evidence in prior research [85–88], which showcases that different models perform better depending on the characteristics of the independent and dependent variables in the data. When using File Data and employing the single model method, the best performance was achieved using CatBoost in 13 (of 16) fields. This result can be interpreted in light of previous studies [75,89], and suggests the specialized performance of CatBoost in handling categorical data. Additionally, according to the MSE and RMSE metrics, the File Data generally demonstrated superior performance compared with OpenAPI Data. We also observed only a relatively small deviation of the RMSE metrics within each field when using File Data; nevertheless, when using OpenAPI Data, there were significant differences in the performance metrics across fields. For example, when applying the random forest model trained on integrated data in the Public Administration field using File Data, the difference between the RMSE maximum value of 0.745 and the minimum value of 0.600 was 0.145; this was smaller than the difference of 3.329 for the same conditions when using OpenAPI Data.

Regarding the implications of distinguishing open data by field, the performance of the models trained with integrated data and field-specific data differed because of variations in the quantity of

 doi:10.20944/preprints202406.1319.v1

accumulated data and field-specific metadata among the 16 fields. For File Data, better performance was generally achieved when the model was trained with integrated data—with the exceptions of the Science & Technology and Finance fields. For OpenAPI Data, a superior performance was observed when the model was trained with integrated data in eight fields (Education, Transportation & Logistics, Agriculture & Fisheries, Culture & Tourism, Law, Healthcare, Industry & Employment, Environment & Meteorology); in the other eight fields (Public Administration, Science & Technology, Land Management, Social Welfare, Food & Health, Disaster & Safety, Finance, Unification & Diplomacy), models trained on field-specific data showed superior performance. These research findings suggest that OpenAPI Data, which are often utilized in real-time and continuous services, can more effectively reflect the characteristics of field-specific data compared with File Data—and this is corroborated by past research [90]. If this understanding is applied to open data openness and utilization in practical, real-world scenarios, and the approach we propose for pre-evaluating and diagnosing open data utilization is implemented, it may help address the ongoing garbage data issues [91,92] related to open data. These findings are expected to serve as a catalyst for accelerating the process of unveiling the details of open data utilization.

6. Conclusion

This study applied machine learning methods to propose an alternative approach for the proactive quantification and evaluation of open data utilization. Regarding academic significance, this research empirically confirmed that building multiple machine learning models and comparing their performance is useful to measure the intangible value of open government data utilization. In so doing, the study overcomes the limitations of previous related studies and expands the horizon of open data utilization measurement, delivering a novel alternative methodology for such procedures. Additionally, this study delivers evidence showing that it is appropriate to consider the attributes of open data (fields and utilization methods) when deciding on which algorithm to apply and training data to use for machine learning approaches.

Regarding practical significance, the proposed approach can be applied in efforts to increase real-world open data utilization. Specifically, its use may enable stakeholders to accurately pinpoint, in advance, the data that they need to disclose to secure high-usability for the open data that is made available. From the perspective of consumers, the tool supporting the provision of highly-usable open data may then help with the creation of various business opportunities.

Regarding policy implications, the alternative approach proposed in this study may allow for a policy focus shift. In particular, while current open data policies focus on “quantity expansion,” the assessment of open data utilization before data provision may make possible a greater focus on “quality enhancement” in related policies. We also suggest for those involved to consider these findings in light of a comprehensive consideration of the indirect factors (law, governance, policy, and technology) emphasized in prior research to influence open data utilization.

Regarding limitations, the machine learning algorithms utilized in this study are tree-based algorithms, and thus other algorithm types were not examined. In addition, for the target indicators of open data utilization, we have established and utilized normalized indicators based on the number of downloads for File Data and the number of applications and calls for OpenAPI Data; there are limitations in the use of these indicators for qualitative analyses focused on tracking and understanding how open data is being used in businesses from an outcome perspective. This limitation may be addressed in the future if digital rights management is applied to open data, as it may enable open data utilization tracking [93]. Furthermore, since we used the official classification system for the public sector in South Korea to classify the 16 fields related to open data utilization, the scalability of the proposed method is limited. Further research should to consider open data attributes beyond those analyzed in this study.

Future research directions include expanding the scope of investigations unstructured data (i.e., heavily utilized in high-level artificial intelligence businesses), so that we can predict the utilization of unstructured data. Convergence analyses with data actually containing different open data attributes could also help us overcome the limitations inherent in the metadata used in this study.

 doi:10.20944/preprints202406.1319.v1

Finally, researchers are recommended to probe into the relationship between post- and indirect evaluation factors associated with open data utilization.

Author Contributions: Conceptualization, J.J. and K.C.; Methodology, J.J. and K.C.; Software, J.J. and K.C.; Validation, J.J. and K.C.; Formal analysis, J.J. and K.C.; Writing—original draft preparation, J.J. and K.C.; Writing—review and editing, J.J. and K.C.; Supervision, K.C. All authors have read and agreed to the published version of the manuscript.

Funding: This study received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are provided and managed by the South Korean government in the Open Government Data portal (data.go.kr).

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

References

1. Gabryelczyk, R., Has COVID-19 accelerated digital transformation? Initial lessons learned for public administrations. Information Systems Management **2020**, 37, (4), 303-309.
2. Hamari, J.; Sjöklint, M.; Ukkonen, A., The sharing economy: Why people participate in collaborative consumption. Journal of the association for information science and technology **2016**, 67, (9), 2047-2059.
3. Niankara, I. In Sustainability through open data sharing and reuse in the digital economy, 2022 International Arab Conference on Information Technology (ACIT), **2022**; pp 1-11.
4. Helbig, R.; von Höveling, S.; Solsbach, A.; Marx Gómez, J., Strategic analysis of providing corporate sustainability open data. Intelligent Systems in Accounting, Finance and Management **2021**, 28, (3), 195-214.
5. Peled, A., When transparency and collaboration collide: The USA open data program. Journal of the American society for information science and technology **2011**, 62, (11), 2085-2094
6. O'Hara, K. In Transparency, open data and trust in government: shaping the infosphere, Proceedings of the 4th annual ACM web science conference, 2012; pp 223-232.
7. Lnenicka, M.; Nikiforova, A., Transparency-by-design: What is the role of open data portals? Telematics and Informatics **2021**, 61, 101605.
8. Hong, Y., A Study on Policies for Activating the Use of Public Data. Journal of the Korean Data & Information Science Society, **2014**, 25, 4, 769-777.
9. Janssen, M.; Charalabidis, Y.; Zuiderwijk, A., Benefits, adoption barriers and myths of open data and open government. Information systems management, **2012**, 29, (4), 258-268.
10. Weerakkody, V.; Irani, Z.; Kapoor, K.; Sivarajah, U.; Dwivedi, Y. K., Open data and its usability: an empirical view from the Citizen's perspective. Information Systems Frontiers **2017**, 19, 285-300.
11. Go, K., Study on Value Creation Strategies of Public Data. Proceedings of the Korean Association of Public Administration 2018, **2018**, 3473-3491.
12. Yoon, S. O.; Hyun, J. W., A Study on the Current Status Analysis and Improvement Measures of Public Data Opening Policies: Focusing on the Case of National Priority Data Opening in the Public Data Portal. Korean Journal of Public Administration **2019**, 33, (1), 219-247.
13. Kim, Eun-Seon, A Study on Legal System Improvement Measures for Promoting the Openness and Utilization of Public Data - Focusing on Cases of Refusal to Provide Public Data. Information Policy **2023**, 30, (2), 46-67.
14. Kim, Min-Ho; Lee, Bo-Oak, Trends and Implications of the Revision of the EU Directive on Public Open Data. Sungkyunkwan Law Review **2020**, 32, (1), 1-30.
15. Devins, C.; Felin, T.; Kauffman, S.; Koppl, R., The law and big data. Cornell JL & Public Policy **2017**, 27, 357.
16. Tan, E., Designing an AI compatible open government data ecosystem for public governance. **2022**.
17. Kim Geun-Hyun, Jung Seong-Hoon, Yang Jae-Dong, and Wi Joo-Yeon, A Policy Study on Public Data for the Past 10 Years Using Big Data Analysis Techniques: Focusing on Comparative Analysis by Administration. National Policy Research **2023**, 37, (4), 45-67.
18. Jetzek, T.; Avital, M.; Bjørn-Andersen, N. In Generating Value from Open Government Data, ICIS, 2013; **2013**.
19. Osagie, E.; Waqar, M.; Adebayo, S.; Stasiewicz, A.; Porwol, L.; Ojo, A. In Usability evaluation of an open data platform, Proceedings of the 18th annual international conference on digital government research, 2017; **2017**; pp 495-504.

20. Máchová, R.; Volejníková, J.; Lněnička, M., Impact of e-government development on the level of corruption: Measuring the effects of related indices in time and dimensions. Review of Economic Perspectives **2018**, 18, (2), 99-121.
21. Khurshid, M. M.; Zakaria, N. H.; Rashid, A.; Shafique, M. N., Examining the factors of open government data usability from academician's perspective. International Journal of Information Technology Project Management (IJITPM) **2018**, 9, (3), 72-85.
22. Hagen, L.; Keller, T. E.; Yerden, X.; Luna-Reyes, L. F., Open data visualizations and analytics as tools for policy-making. Government Information Quarterly **2019**, 36, (4), 101387.
23. Joseph, A., Schumpeter, The theory of economic development: An inquiry into profits, capital, credit, interest, and the business cycle. **1934**.
24. Bason, C., Leading public sector innovation. Bristol: Policy Press: **2010**; Vol. 10.
25. Zuiderwijk, A.; Janssen, M.; Davis, C., Innovation with open data: Essential elements of open data ecosystems. Information polity **2014**, 19, (1-2), 17-33.
26. Blakemore, M.; Craglia, M., Access to public-sector information in Europe: Policy, rights, and obligations. The Information Society **2006**, 22, (1), 13-24.
27. Charalabidis, Y.; Zuiderwijk, A.; Alexopoulos, C.; Janssen, M.; Höchtl, J.; Ferro, E., The world of open data. Public Administration and Information Technology. Cham: Springer International Publishing. doi **2018**, 978-3.
28. European Commission, "Digital agenda: Turning government data into gold, European Commission, Brussels", **2011**.
29. Zhang, J.; Dawes, S. S.; Sarkis, J., Exploring stakeholders' expectations of the benefits and barriers of e-government knowledge sharing. Journal of Enterprise Information Management **2005**, 18, (5), 548-567.
30. Kitsios, F.; Papachristos, N.; Kamariotou, M. In Business models for open data ecosystem: Challenges and motivations for entrepreneurship and innovation, 2017 IEEE 19th Conference on Business Informatics (CBI), 2017; IEEE: **2017**; pp 398-407.
31. European Commission, "Digital agenda: Commission's open data strategy, questions & answers", **2013**.
32. Ministry of the Interior and Safety, "2021 Administrative Safety White Paper", **2022**.
33. Janssen, M.; Zuiderwijk, A., Infomediary business models for connecting open data providers and users. Social Science Computer Review **2014**, 32, (5), 694-711.
34. Borgesius, F. Z.; Gray, J.; Van Eechoud, M., Open data, privacy, and fair information principles: Towards a balancing framework. Berkeley Technology Law Journal **2015**, 30, (3), 2073-2131.
35. Thompson, N.; Ravindran, R.; Nicosia, S., Government data does not mean data governance: Lessons learned from a public sector application audit. Government information quarterly **2015**, 32, (3), 316-322.
36. Zuiderwijk, A.; Janssen, M., Open data policies, their implementation and impact: A framework for comparison. Government information quarterly **2014**, 31, (1), 17-29.
37. Bertot, J. C.; Gorham, U.; Jaeger, P. T.; Sarin, L. C.; Choi, H., Big data, open government and e-government: Issues, policies and recommendations. Information polity **2014**, 19, (1-2), 5-16.
38. Máchová, R.; Lněnička, M., Evaluating the quality of open data portals on the national level. Journal of theoretical and applied electronic commerce research **2017**, 12, (1), 21-41.
39. Osagie, E.; Waqar, M.; Adebayo, S.; Stasiewicz, A.; Porwol, L.; Ojo, A. In Usability evaluation of an open data platform, Proceedings of the 18th annual international conference on digital government research, 2017; **2017**; pp 495-504.
40. Vetrò, A.; Canova, L.; Torchiano, M.; Minotas, C. O.; Iemma, R.; Morando, F., Open data quality measurement framework: Definition and application to Open Government Data. Government Information Quarterly **2016**, 33, (2), 325-337.
41. Kourou, K.; Exarchos, T. P.; Exarchos, K. P.; Karamouzis, M. V.; Fotiadis, D. I., Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal **2015**, 13, 8-17.
42. Kim, E.; Lee, Y.; Choi, J.; Yoo, B.; Chae, K. J.; Lee, C. H., Machine Learning-based Prediction of Relative Regional Air Volume Change from Healthy Human Lung CTs. KSII Trans. Internet Inf. Syst. **2023**, 17, (2), 576-590.
43. Kruppa, J.; Ziegler, A.; König, I. R., Risk estimation and risk prediction using machine-learning methods. Human genetics **2012**, 131, 1639-1654.
44. Xayasouk, T.; Lee, H.; Lee, G., Air pollution prediction using long short-term memory (LSTM) and deep autoencoder (DAE) models. Sustainability **2020**, 12, (6), 2570.
45. Mosavi, A.; Ozturk, P.; Chau, K.-w., Flood prediction using machine learning models: Literature review. Water **2018**, 10, (11), 1536.
46. Ahmed, A. N.; Othman, F. B.; Afan, H. A.; Ibrahim, R. K.; Fai, C. M.; Hossain, M. S.; Ehteram, M.; Elshafie, A., Machine learning methods for better water quality prediction. Journal of Hydrology **2019**, 578, 124084.

 doi:10.20944/preprints202406.1319.v1

47. Lee, D.-S.; Choi, W. I.; Nam, Y.; Park, Y.-S., Predicting potential occurrence of pine wilt disease based on environmental factors in South Korea using machine learning algorithms. Ecological informatics **2021**, 64, 101378.
48. Zhang, L.; Wen, J.; Li, Y.; Chen, J.; Ye, Y.; Fu, Y.; Livingood, W., A review of machine learning in building load prediction. Applied Energy **2021**, 285, 116452.
49. Paltrinieri, N.; Comfort, L.; Reniers, G., Learning about risk: Machine learning for risk assessment. Safety science **2019**, 118, 475-486.
50. Hegde, J.; Rokseth, B., Applications of machine learning methods for engineering risk assessment–A review. Safety science **2020**, 122, 104492.
51. Ahuja, G.; Morris Lampert, C., Entrepreneurship in the large corporation: A longitudinal study of how established firms create breakthrough inventions. Strategic management journal **2001**, 22, (6-7), 521-543.
52. Wu, J.-L.; Chang, P.-C.; Tsao, C.-C.; Fan, C.-Y., A patent quality analysis and classification system using self-organizing maps with support vector machine. Applied soft computing **2016**, 41, 305-316.
53. Cho Hyunjin; Lee Hakyun, Patent Quality Prediction Using Machine Learning Techniques. Proceedings of the Korean Institute of Industrial Engineers Spring Conference **2018**, 1343-1350.
54. Erdogan, Z.; Altuntas, S.; Dereli, T., Predicting patent quality based on machine learning approach. IEEE Transactions on Engineering Management **2022**.
55. Kim, K.; Hong, J.-s., A hybrid decision tree algorithm for mixed numeric and categorical data in regression analysis. Pattern Recognition Letters **2017**, 98, 39-45.
56. Cha, G.-W.; Moon, H.-J.; Kim, Y.-C., Comparison of random forest and gradient boosting machine models for predicting demolition waste based on small datasets and categorical variables. International Journal of Environmental Research and Public Health **2021**, 18, (16), 8530.
57. Foody, G. M.; Mathur, A.; Sanchez-Hernandez, C.; Boyd, D. S., Training set size requirements for the classification of a specific class. Remote Sensing of Environment **2006**, 104, (1), 1-14.
58. Ramezan, C. A.; Warner, T. A.; Maxwell, A. E.; Price, B. S., Effects of training set size on supervised machine-learning land-cover classification of large-area high-resolution remotely sensed data. Remote Sensing **2021**, 13, (3), 368.
59. Ahmad, I.; Basheri, M.; Iqbal, M. J.; Rahim, A., Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection. IEEE access **2018**, 6, 33789-33795.
60. Daghistani, T.; Alshammari, R., Comparison of statistical logistic regression and random forest machine learning techniques in predicting diabetes. Journal of Advances in Information Technology Vol **2020**, 11, (2), 78-83.
61. Suenaga, D.; Takase, Y.; Abe, T.; Orita, G.; Ando, S. In Prediction accuracy of Random Forest, XGBoost, LightGBM, and artificial neural network for shear resistance of post-installed anchors, Structures, 2023; Elsevier: **2023**; pp 1252-1263.
62. Shehadeh, A.; Alshboul, O.; Al Mamlook, R. E.; Hamedat, O., Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression. Automation in Construction **2021**, 129, 103827.
63. Muslim, M. A.; Dasril, Y., Company bankruptcy prediction framework based on the most influential features using XGBoost and stacking ensemble learning. International Journal of Electrical and Computer Engineering (IJECE) **2021**, 11, (6), 5549-5557.
64. Rebala, G.; Ravi, A.; Churiwala, S.; Rebala, G.; Ravi, A.; Churiwala, S., Machine learning definition and basics. An introduction to machine learning **2019**, 1-17.
65. West, S. G.; Finch, J. F.; Curran, P. J., Structural equation models with nonnormal variables: Problems and remedies. **1995**.
66. Hong, S.; Malik, M. L.; Lee, M.-K., Testing configural, metric, scalar, and latent mean invariance across genders in sociotropy and autonomy using a non-Western sample. Educational and psychological measurement **2003**, 63, (4), 636-654.
67. Kwon, H.; Park, J.; Lee, Y., Stacking ensemble technique for classifying breast cancer. Healthcare informatics research **2019**, 25, (4), 283.
68. Painsky, A.; Rosset, S.; Feder, M., Large alphabet source coding using independent component analysis. IEEE Transactions on Information Theory **2017**, 63, (10), 6514-6529.
69. Kim, K.; Hong, J.-s., A hybrid decision tree algorithm for mixed numeric and categorical data in regression analysis. Pattern Recognition Letters **2017**, 98, 39-45.
70. Qin, X.; Han, J., Variable selection issues in tree-based regression models. Transportation Research Record **2008**, 2061, (1), 30-38.
71. Géron, A., Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. " O'Reilly Media, Inc.": **2022**.
72. Dangeti, P., Statistics for machine learning. Packt Publishing Ltd.: **2017**.
73. Chen, T.; Guestrin, C. In Xgboost: A scalable tree boosting system, Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016; **2016**; pp 785-794.

 doi:10.20944/preprints202406.1319.v1

74. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y., Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems **2017**, 30.
75. Hancock, J. T.; Khoshgoftaar, T. M., CatBoost for big data: an interdisciplinary review. Journal of big data **2020**, 7, (1), 94.
76. Wei, X.; Rao, C.; Xiao, X.; Chen, L.; Goh, M., Risk assessment of cardiovascular disease based on SOLSSA-CatBoost model. Expert Systems with Applications **2023**, 219, 119648.
77. Jabeur, S. B.; Gharib, C.; Mefteh-Wali, S.; Arfi, W. B., CatBoost model and artificial intelligence techniques for corporate failure prediction. Technological Forecasting and Social Change **2021**, 166, 120658.
78. Luo, M.; Wang, Y.; Xie, Y.; Zhou, L.; Qiao, J.; Qiu, S.; Sun, Y., Combination of feature selection and catboost for prediction: The first application to the estimation of aboveground biomass. Forests **2021**, 12, (2), 216.
79. Jin, Y.; Ye, X.; Ye, Q.; Wang, T.; Cheng, J.; Yan, X., Demand forecasting of online car-hailing with stacking ensemble learning approach and large-scale datasets. IEEE Access **2020**, 8, 199513-199522.
80. Acquah, J.; Owusu, D. K.; Anafo, A., Application of Stacked Ensemble Techniques for Classifying Recurrent Head and Neck Squamous Cell Carcinoma Prognosis. Asian Journal of Research in Computer Science **2024**, 17, (4), 77-94.
81. Sahin, E. K.; Demir, S., Greedy-AutoML: A novel greedy-based stacking ensemble learning framework for assessing soil liquefaction potential. Engineering Applications of Artificial Intelligence **2023**, 119, 105732.
82. Aswin, S.; Geetha, P.; Vinayakumar, R. In Deep learning models for the prediction of rainfall, 2018 International Conference on Communication and Signal Processing (ICCSP), **2018**; IEEE: 2018; pp 0657-0661.
83. Almalaq, A.; Edwards, G. In A review of deep learning methods applied on load forecasting, 2017 16th IEEE international conference on machine learning and applications (ICMLA), **2017**; IEEE: 2017; pp 511-516.
84. Si, B.; Ni, Z.; Xu, J.; Li, Y.; Liu, F., Interactive effects of hyperparameter optimization techniques and data characteristics on the performance of machine learning algorithms for building energy metamodeling. Case Studies in Thermal Engineering **2024**, 104124.
85. Satoła, A.; Satoła, K., Performance comparison of machine learning models used for predicting subclinical mastitis in dairy cows: bagging, boosting, stacking and super-learner ensembles versus single machine learning models. Journal of Dairy Science **2024**.
86. Zhang, H.; Zhu, T., Stacking Model for Photovoltaic-Power-Generation Prediction. Sustainability **2022**, 14, (9), 5669.
87. Park, U.; Kang, Y.; Lee, H.; Yun, S., A stacking heterogeneous ensemble learning method for the prediction of building construction project costs. Applied sciences **2022**, 12, (19), 9729.
88. Um Hanuel; Kim Jaesung; Choi Sangok, Verification of Machine Learning-Based Corporate Bankruptcy Risk Prediction Model and Policy Suggestions: Focused on Improvement through Stacking Ensemble Model. Journal of Intelligence and Information Systems Research **2020**, 26(2), 105-129.
89. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A. V.; Gulin, A., CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems **2018**, 31.
90. Sugiyama, M., Introduction to statistical machine learning. Morgan Kaufmann: **2015**.
91. Grimes, D. A., Epidemiologic research using administrative databases: garbage in, garbage out. Obstetrics & Gynecology **2010**, 116, (5), 1018-1019.
92. Kilkenny, M. F.; Robinson, K. M., Data quality:"Garbage in–garbage out". In SAGE Publications Sage UK: London, England: **2018**; Vol. 47, pp 103-105.
93. Hartung, F.; Ramme, F., Digital rights management and watermarking of multimedia content for m-commerce applications. IEEE communications magazine **2000**, 38, (11), 78-84.