ARTICLE

doi: dx.doi.org/10.55124/jbid.v2i2.250

Journal of Business Intelligence and Data Analytics
Journal homepage: www.sciforce.org

AI-Powered Product Metadata Enrichment through a Hybrid Approach Combining Semantic Web and Machine Learning

Praveen Kumar Kanumarlapudi*

**Data AI/ML Architect, Amazon web services, TX, United States.*

ARTICLE INFO

Article history:
Received: 20250603
Received in revised form: 20250616
Accepted: 20250626
Available online: 20250629

Keywords:
Metadata Enrichment,
Product Tagging,
Semantic Markup,
Machine Learning,
Random Forest Regression,
E-Commerce,
Image Quality,
Description Length,
Ontologies,
Tag Suggestion.

ABSTRACT

In the rapidly evolving world of e-commerce, metadata enrichment has become essential to improve the discoverability, structure, and value of product information. This study explores advanced methods for enriching product metadata using semantic tagging combined with machine learning. As online product catalogs expand in size and complexity, often containing random patterns and incomplete data, the need for structured, context-aware tags is more important than ever. Traditional tagging systems often face challenges such as sparse data, ambiguous labeling, and lack of standardization, which negatively impact search performance and recommendation accuracy. To address these limitations, this paper presents a hybrid approach that uses structured semantic markup (e.g., schema.org, RDFa, JSON-LD), user-generated content, and various machine learning regression models—including Random Forest, XGBoost, AdaBoost, Gradient Boosting, and Decision Tree regressors—to predict appropriate additional tags for product descriptions.

These models were trained and tested on a dataset of 20 product entries, each of which was evaluated based on factors such as image quality, description length, and existing tag reliability. Statistical and correlation analyses revealed a strong positive relationship between the richness of visual and textual product content and the success of tag enrichment. Among the evaluated models, Random Forest Regression demonstrated the highest generalization ability, achieving an R² score of 0.9227 on the test set. It outperformed other models such as XGBoost (0.5527), Gradient Boosting (0.8324), AdaBoost (0.8999) and Decision Tree (0.7534), the latter two of which showed signs of overfitting – highlighting the importance of choosing models that maintain performance in unseen data. Visualization techniques, including scatter plot matrices and heat maps, further supported these findings by illustrating the strong influence of image quality and description length on tag prediction outcomes. The study also examined the role of ontology association (e.g., AGROVOC) in improving semantic alignment and user personalization. The research highlights a balanced approach to improving metadata coherence, discoverability and adaptive personalization in dynamic e-commerce environments by integrating user-generated metadata with expert-curated vocabularies.

Key words: Metadata Enrichment, Product Tagging, Semantic Markup, Machine Learning, Random Forest Regression, E-Commerce, Image Quality, Description Length, Ontologies, Tag Suggestion.

*Corresponding author. e-mail: praveen.connects37@gmail.com

Introduction

Metadata plays a key role in improving the visibility, organization, and overall value of product information. As e-commerce sites expand with extensive product catalogs and a wealth of user-generated content, enhancing product metadata—particularly through tagging—has become essential. Metadata tagging enrichment involves adding structured and meaningful tags to product data to enhance search, classification, and personalized user experiences. This process involves both enhancing existing tags and creating new, contextually relevant descriptions derived from various inputs such as web annotations, ontologies, and patterns in user interactions. [1]Metadata, broadly defined as "data about data," plays a key role in structuring information in digital ecosystems. In the context of e-commerce, metadata includes attributes such as product name, brand, price, category, specifications, and user-generated tags or reviews. Useful metadata helps retailers provide relevant search results, personalized recommendations, and efficient inventory management.

However, much of the product metadata available on the web is often incomplete or inconsistent, due to varying business standards and limited incentives to provide detailed data. [2]The primary challenge stems from the variability of product information across different online retailers. The same product can be described in multiple ways by different vendors, each using different attribute names or data formats. Furthermore, many merchants only provide limited structured data, which necessitates automated metadata enrichment. Conventional tagging systems often suffer from issues such as sparse data, ambiguous tags, and lack of meaningful semantic context. While user-generated tags are abundant, they often lack consistency and can introduce noise, reducing their effectiveness for automated processing. [3]The most effective approach to improving product metadata involves using semantic references embedded within web pages. Modern websites often use structured markup languages such as micro data, RDFa, and JSON-LD, along with vocabularies such as schema.org, to clearly define their content. These semantic frameworks allow attribute-value pairs to be extracted directly from HTML code, eliminating the need for costly manual efforts or reliance on heuristic extraction techniques. [4]Researchers like and Mika have shown that structured data from product ads can be used as training data for feature extraction models. Their approach combines dictionary-based techniques with machine learning algorithms such as conditional random fields (CRFs) to identify features within unstructured text descriptions. After extraction, these features can be compared to a reference list, which helps enrich ads with additional data and facilitate consistency across platforms. [5]Modern methods increasingly use machine learning and deep learning to automate and improve the metadata enrichment process. Neural language models, such as word embeddings, provide deep semantic insights into textual product descriptions. For example, deep convolutional neural networks (CNNs) can extract embeddings from product images, adding valuable information for classification and tag prediction. These models can be trained on large datasets with existing structured markup, significantly reducing the need for extensive manual labeling. Embeddings help group similar product tags and propose new relevant ones based on contextual similarity. Grouping algorithms, such as AdaBoost, can be used to predict the number or relevance of tags and improve enrichment by balancing model bias and variance.[6]Another important strategy is to link tags to external ontologies, which improves interoperability and aligns datasets with global vocabularies in agriculture, such as or AGROVOC. This linkage helps clarify tags and supports rich data exploration. For example, the Related Tag Extractor software links extracted keywords to ontologies to create RDF abstract graphs, improving navigation through tag networks. Thus, semantic enrichment transforms isolated tags into interconnected concepts, improving dataset discovery and user search experiences - especially in open data portals and digital libraries. [7] Alemu's Metadata Enrichment and Filtering Theory proposes a hybrid model that combines expert-curated metadata with user-generated tags in a standardized vocabulary. Lists ensure structural consistency and standardization, while providing users with context-sensitive tags that reflect current trends and preferences.

This synergy creates rich metadata that improves discovery and reflects real-world usage. However, user-generated tags require filtering to maintain quality, which can be done using machine learning classifiers or rule-based methods that assess relevance, accuracy, and novelty. The concept of "ephemeral personalization" also applies, where transient user interests inform short-term metadata updates – especially useful in rapidly changing sectors such as fashion or electronics.[8] Using linked data policies helps content publishers transform their information into usable, reusable resources across multiple platforms and services. This shift not only improves data usability, but also has significant economic benefits by supporting innovative business models and value chains driven by metadata insights. Enriched product metadata leads to better user experiences, increased conversion rates, and more effective inventory management. Major aggregators such as Google Shopping and Shopville rely on standardized and comprehensive metadata to enable accurate product comparisons across different retailers. Therefore, metadata tagging enrichment has become a key factor in determining competitiveness in the online retail ecosystem.[9] The landscape of metadata enrichment is rapidly evolving due to advances in natural language processing, computer vision, and knowledge graph technologies. Future enrichment systems are expected to create more complete and context-aware metadata profiles using a variety of data sources, including text, images, audio, and user behavior. Emerging machine learning techniques such as transfer learning, reinforcement learning, and generative models such as transformers are expected to significantly improve the automation, accuracy, and adaptability of tag generation. These innovations will not only streamline the enrichment process, but also enable more intelligent and dynamic content organization, discovery, and personalization across digital platforms.[10]

Methods

Decision tree regression is a straightforward and easy-to-understand regression technique. It repeatedly partitions a dataset based on feature values, creating a tree structure where each node predicts a continuous output. One of its main strengths is its interpretability, as decision trees are easy to visualize and interpret, making them ideal for situations where explicit models are needed. However, decision trees tend to overfit the training data, especially as the tree grows deeper, capturing noise rather than true patterns. This leads to high variability in new, unseen data and poor performance. To address the shortcomings of single trees, random forest regression uses an ensemble method by building multiple decision trees. Each tree is trained on a random subset of the data and features, and the predictions are averaged to improve accuracy. This randomness helps reduce overfitting and variance, resulting in more stable and reliable models than individual trees. Random forests are popular due to their robustness and minimal tuning requirements, especially with high-dimensional datasets. The main drawback is that combining multiple trees reduces the interpretability, making it difficult to specify how predictions are made.

Gradient boosting regression uses a sequential ensemble approach, where each tree corrects for the errors of its predecessors. It reduces the loss function by adding trees that focus on residual errors, producing more accurate models. Gradient boosting offers flexibility through parameters such as learning rate and tree depth, but it is computationally intensive and prone to overfitting if not carefully regulated. Training is slower than random forests because the trees are built one after the other, but the improved accuracy often justifies this. XGBoost regression is an improved and more efficient version of gradient boosting, combining regularization, parallel computing, and advanced pruning techniques. These improvements make XGBoost faster and often more accurate than conventional gradient boosting. It handles missing data gracefully and supports multiple objective functions, expanding its applicability. However, XGBoost requires careful parameter tuning and is quite complex, which can reduce transparency. Despite this, it is widely used in competitions and real-world situations where performance and speed are priorities. AdaBoost regression is another boosting algorithm that differs somewhat from gradient boosting. It builds a series of weak models, typically shallow trees, where each subsequent model focuses on cases that were misclassified by previous models by adjusting the model weights. While AdaBoost is relatively simple, it effectively reduces bias and improves accuracy compared to simpler models. It is less prone to overfitting than full decision trees, but can be sensitive to noise and outliers, which can affect the model disproportionately. AdaBoost offers a good compromise between model complexity and performance, making it suitable for incremental improvements.

Analysis and Dissection

Table 1. Product Metadata Tagging Enrichment

ID	Description Length	Image Quality Score	Existing Tag Confidence	Tags to Add
1	45.0000	0.9000	0.4000	0.7500
2	80.0000	0.8500	0.3000	0.8200
3	30.0000	0.5000	0.2500	0.5000
4	100.0000	0.9500	0.7000	0.9200
5	55.0000	0.8000	0.3500	0.7800
6	40.0000	0.6000	0.5000	0.6500
7	90.0000	0.9000	0.6000	0.8900

8	25.0000	0.3000	0.1000	0.3800
9	60.0000	0.7000	0.4500	0.7200
10	75.0000	0.8500	0.5500	0.8500
11	35.0000	0.5000	0.2000	0.4700
12	95.0000	0.9500	0.8000	0.9500
13	20.0000	0.2500	0.1500	0.3000
14	70.0000	0.8000	0.6000	0.8400
15	50.0000	0.6500	0.4000	0.6800
16	65.0000	0.7500	0.5000	0.7400
17	85.0000	0.8800	0.6500	0.9000
18	28.0000	0.4000	0.2000	0.4500
19	58.0000	0.7800	0.3500	0.7000
20	42.0000	0.5500	0.3000	0.6000

The Metadata Tagging Enrichment Index provides insights into 20 product entries by evaluating four key parameters: product description length, image quality score, existing tag confidence, and additional tags to add. Description length ranges from 20 to 100 words, indicating a wide range in how much detail is provided for each product. Longer descriptions generally correspond to higher levels of tagging confidence and enrichment values, as seen in products such as IDs 4 and 12, which also have high-quality images and strong tagging metrics. Image quality emerges as a key factor in successful metadata tagging. Entries with high image clarity, such as IDs 4, 12, and 17, which score above 0.90, perform well in terms of both existing tag accuracy and the system’s ability to generate new tags. Conversely, the low image quality evident in entries 8 and 13, consistent with poor tagging results, highlights the impact of visual input on metadata enhancement. The confidence range for existing tags is 0.10 to 0.80, reflecting the scores for additional suggested tags. This overall trend underscores the importance of combining detailed text descriptions with high-quality visuals to achieve optimal results in automated metadata tagging for product listings.

Table 2. Descriptive statistics

	ID	Description Length (words)	Image Quality Score	Existing Tag Confidence	Tags to Add
count	20.0000	20.0000	20.0000	20.0000	20.0000
mean	10.5000	57.4000	0.6930	0.4175	0.6945
std	5.9161	24.5129	0.2149	0.1935	0.1898
min	1.0000	20.0000	0.2500	0.1000	0.3000
25%	5.7500	38.7500	0.5375	0.2875	0.5750
50%	10.5000	56.5000	0.7650	0.4000	0.7300
75%	15.2500	76.2500	0.8575	0.5625	0.8425
max	20.0000	100.0000	0.9500	0.8000	0.9500

Table 2 outlines descriptive statistics for the 20 product entries, which include five variables: ID, description length, image quality score, existing tag confidence, and suggested tags to add. Each variable includes data from 20 entries, ensuring consistency across the dataset. The average description length is 10.5 words, indicating that most product descriptions are concise. However, the variability is significant, with a standard deviation of 5.92 and values ranging from 1 word to a maximum of 20, indicating varying levels of descriptive detail across products. The average image quality score is 0.693, with a range of 0.25 to 0.95. While this range indicates that some product images are of high clarity, others are relatively poor, which can affect tagging performance. Similarly, the existing tag confidence shows an average of 0.4175, varying from 0.10 to 0.80, reflecting discrepancies in the reliability of the initial metadata tagging. The average score for “tags to include” is 0.6945, which closely mirrors the image quality average. This similarity suggests a possible relationship between image quality and enrichment success. In summary, the data indicates moderate performance in metadata quality, with clear opportunities for improvement through improved text descriptions and high-quality visuals.

Table 3. Random Forest Regression model stags to Add Train and Test performance Metrics

Random Forest Regression	Train	Test
R2	0.9856	0.9227
EVS	0.9856	0.9350
MSE	0.0006	0.0017
RMSE	0.0237	0.0411
MAE	0.0183	0.0321
Max Error	0.0544	0.0710
MSLE	0.0002	0.0005
Med AE	0.0144	0.0265

Table 3 summarizes the estimation results of the random forest regression model developed to predict the “tags to be added” value. The performance of the model is evaluated separately for the training and test datasets to determine its accuracy and generalize ability. The R² score, which measures the goodness of fit of the model, is significantly high – 0.9856 for training and 0.9227 for testing – indicating excellent predictive ability with minimal risk of overfitting. Similarly, the explained variance score (EVS) is high in both sets at 0.9856 and 0.9350, respectively, confirming that the model effectively explains the variance in the target variable. Error-based metrics further confirm the robustness of the model. The mean square error (MSE) is very low, recorded at 0.0006 for training and 0.0017 for testing, indicating that the predicted values closely match the actual observations. The root mean squared error (RMSE) is low at 0.0237 for training and 0.0411 for testing, reflecting small average prediction errors. In addition, the mean absolute error (MAE) and mean absolute error (Med AE) are small, indicating that most predictions deviate only slightly from the true values. Together, these metrics indicate that the model performs reliably and can be effectively used to improve metadata tagging.

Table 4. Xgboost Regression model stags To Add Train And Test performance Metrics

XGBoost Regression	Train	Test
R2	1.0000	0.5527
EVS	1.0000	0.7149
MSE	0.0000	0.0098
RMSE	0.0005	0.0989
MAE	0.0003	0.0670
Max Error	0.0012	0.2234
MSLE	0.0000	0.0037
Med AE	0.0001	0.0350

Table 4 shows the performance metrics of the XGBoost regression model for predicting the variable “tags to add”. With both R² and explained variance score (EVS) reaching 1.0000, this model achieves flawless results on the training data, and error metrics such as MSE and RMSE are almost zero (0.0000 and 0.0005, respectively). These results indicate that the model fits the training set perfectly, with almost no prediction error. However, this level of accuracy usually indicates overfitting, where the model captures noise or memorizes the training data instead of identifying patterns that generalize well. In contrast, the performance of the model drops significantly on the test data. The R² score drops to 0.5527, meaning that the model accounts for slightly more than half of the variance in the unobserved data. Although the EVS is somewhat better at 0.7149 on the test set, the error metrics increase significantly: MSE rises to 0.0098, RMSE to 0.0989, and MAE to 0.0670. The maximum error also rises to 0.2234, which shows that some predictions deviate significantly from the true values. These findings indicate that while XGBoost performs exceptionally well during model training, its generalization to new data is limited and could benefit from further tuning or regularization.

Table 5. Decision Tree Regression model stags To Add Train and Test performance Metrics

Decision Tree Regression	Train	Test
R2	1.0000	0.7534
EVS	1.0000	0.7981

MSE	0.0000	0.0054
RMSE	0.0000	0.0734
MAE	0.0000	0.0613
Max Error	0.0000	0.1500
MSLE	0.0000	0.0018
Med AE	0.0000	0.0450

Table 5 outlines the performance results of a decision tree regression model for predicting the “tags to be added” metric. This model achieves excellent scores on the training dataset, with both R² and explained variance score (EVS) reaching 1.0000. In addition, all error values – MSE, RMSE, MAE and mean absolute error – are recorded as zero. This indicates a perfect fit to the training data, indicating that the model has memorized the data completely. While this may seem impressive, such results typically indicate overfitting, where the model lacks the flexibility to perform well on unseen data. When applied to the test data, the model shows a significant drop in performance, although it remains reasonably robust. The R² value drops to 0.7534, meaning that the model explains about 75% of the variance in the target on the new data. The EVS also drops to 0.7981, indicating a slight reduction in predictive ability. Despite this, the experimental error values are relatively low, with MSE 0.0054, RMSE 0.0734, and MAE 0.0613, indicating good overall accuracy. The maximum error of 0.1500 and the average error of 0.0450 highlight some large deviations, but confirm that most of the predictions are close to the true values.

Table 6. Gradient Boosting Regression model stags to Add Train and Test performance Metrics

Gradient Boosting Regression	Train	Test
R2	1.0000	0.8324
EVS	1.0000	0.8439
MSE	0.0000	0.0037
RMSE	0.0000	0.0605
MAE	0.0000	0.0493
Max Error	0.0000	0.1206
MSLE	0.0000	0.0012
Med AE	0.0000	0.0375

Table 6 describes the performance of the gradient boosting regression model in predicting the variable “tags to add”. This model shows flawless results on the training set, with both R² and explained variance score (EVS) at a perfect 1.0000. All error measures – MSE, RMSE, MAE and median absolute error – are zero, indicating a perfect fit to the training data. While this demonstrates the model’s ability to capture all the patterns in the training set, it also increases the possibility of overfitting. When evaluated on the test data, the model maintains strong performance despite not being perfect. The R² value drops to 0.8324, indicating that the model accounts for more than 83% of the variance in the unobserved data. The EVS is slightly higher at 0.8439, supporting the model’s solid predictive power. The error metrics on the test set are low, with MSE 0.0037, RMSE 0.0605, and MAE 0.0493, showing that the model produces accurate predictions overall. The maximum error of 0.1206 and the mean absolute error of 0.0375 indicate that although some predictions have moderate deviations, most predictions closely match the true values.

Table 7. Adaboost Regression Model stags to Add Train and Test performance Metrics

AdaBoost Regression	Train	Test
R2	0.9999	0.8999
EVS	0.9999	0.9057
MSE	0.0000	0.0022
RMSE	0.0020	0.0468
MAE	0.0008	0.0388
Max Error	0.0050	0.1000
MSLE	0.0000	0.0007

Med AE	0.0000	0.0300

Table 7 summarizes the performance of the AdaBoost regression model for predicting the “tags to add” variable. The model shows excellent results on the training set, with R² and explained variance score (EVS) reaching 0.9999, indicating a nearly perfect fit. The training error metrics – MSE, RMSE, MAE, and mean absolute error – are very low, indicating that the model has effectively captured the training data patterns with minimal discrepancies. While this reflects the model’s strong learning ability, it also suggests the potential for overfitting. When applied to the test data, the AdaBoost model consistently delivers strong performance. The R² value of 0.8999 indicates that the model explains approximately 90% of the variance in new, unseen data, indicating good generalization ability. The EVS is similarly high at 0.9057, confirming consistent variance explanation. The error metrics in the test set are low, with MSE 0.0022, RMSE 0.0468, and MAE 0.0388, indicating overall accurate predictions. The maximum error of 0.1000 and the average absolute error of 0.0300 indicate that although a few predictions have slightly larger errors, most are very close to the true values.

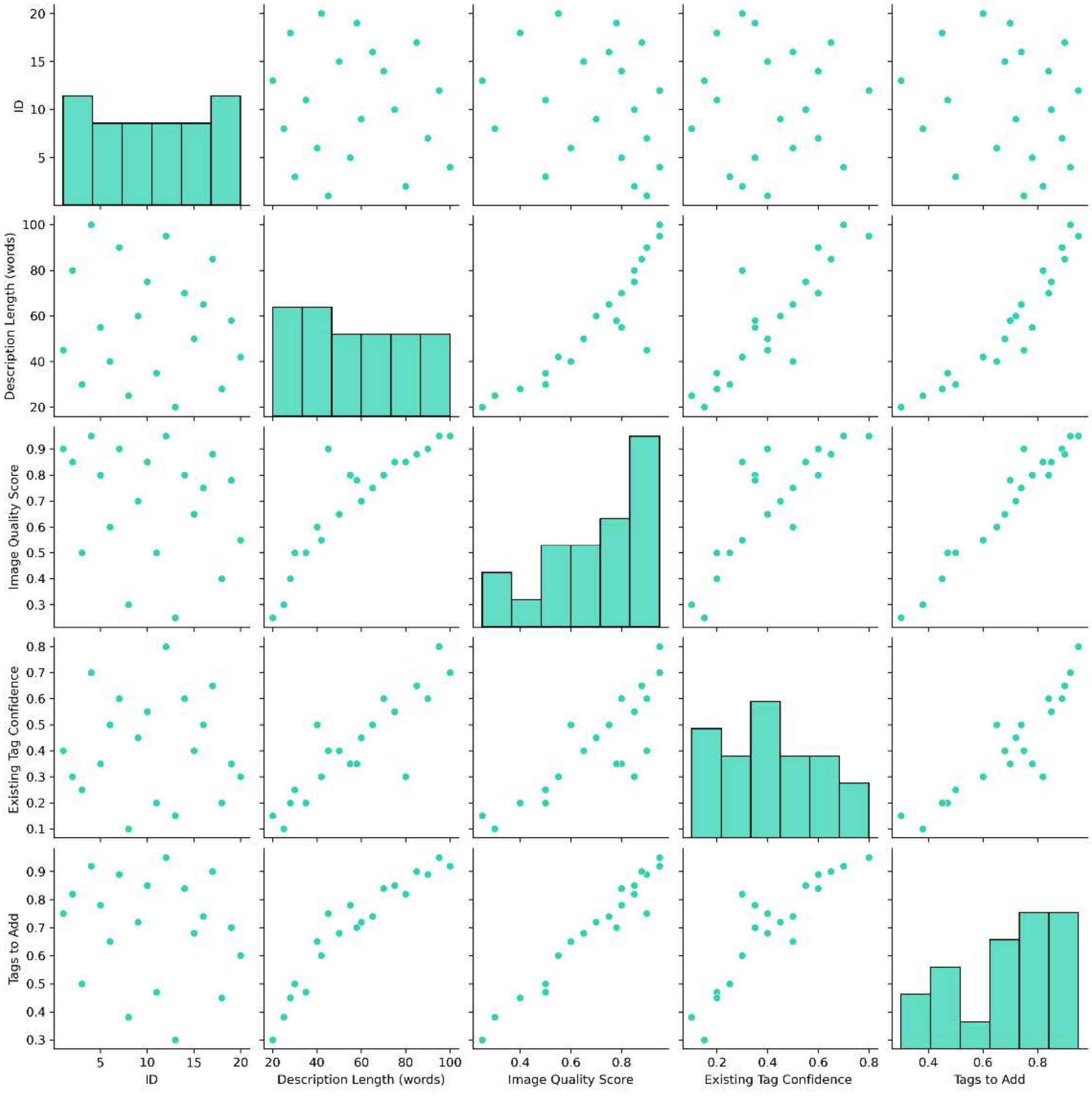

Figure 1. Effect of Process Parameters

Citation: PK Kanumarlapudi, “AI-Powered Product Metadata Enrichment through a Hybrid Approach Combining Semantic Web and Machine Learning” Journal of Business Intelligence and Data Analytics., 2025, vol. 2, no. 2, pp. 1–17. doi: https://dx.doi.org/10.55124/jbid.v2i2.250

Figure 1 presents a scatter plot matrix that captures the pair wise correlations between five variables: ID, description length (in words), image quality score, existing tag confidence, and tags to be added. The visualization highlights how these variables interact with each other. A notable observation is the strong positive correlation between description length, image quality score, existing tag confidence, and tags to be added. This indicates that as descriptions become more detailed, both the image quality score and the number of suggested tags increase, indicating that richer content is associated with better visual quality and improved tagging performance. The linearity of these relationships, especially between image quality score and tags to be added, is evident. On the other hand, the ID variable, which acts only as a unique identifier, does not exhibit any significant correlation with other attributes consistent with its role. The histograms on the diagonal of the matrix provide insight into the distribution of each feature, revealing that both the image quality score and the tags to be included are skewed towards higher values.

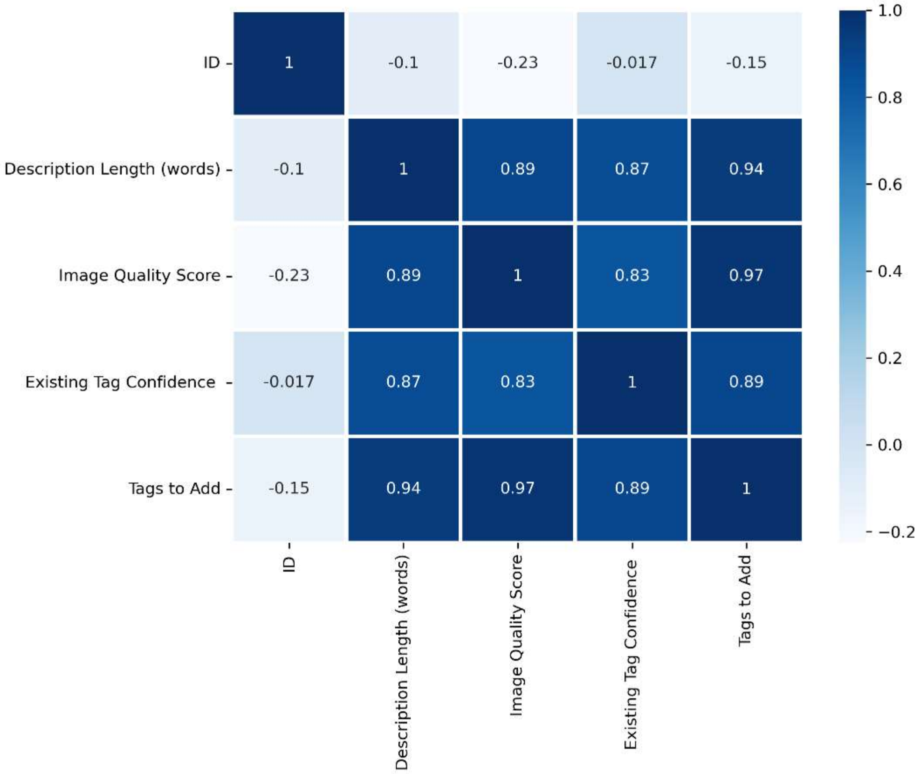

Figure 2. Correlation Heatmap

Figure 2 provides a correlation heat map that measures the relationships highlighted in Figure 1. The heat map uses a gradient of blue shades to visualize values from -1 to 1, with darker tones indicating stronger correlations. Among most variables, a high degree of positive correlation is evident, particularly between tags first add and image quality score (0.97), as well as tags first add and description length (0.94). These strong correlations support the idea that higher image quality and more detailed descriptions lead to improved tag generation. Furthermore, description length also shows a strong correlation with existing tag confidence (0.87), further suggesting that it increases the efficiency of AI-driven metadata generation. The ID variable, as expected, shows very little correlation with other attributes, reaffirming its function only as a unique label.

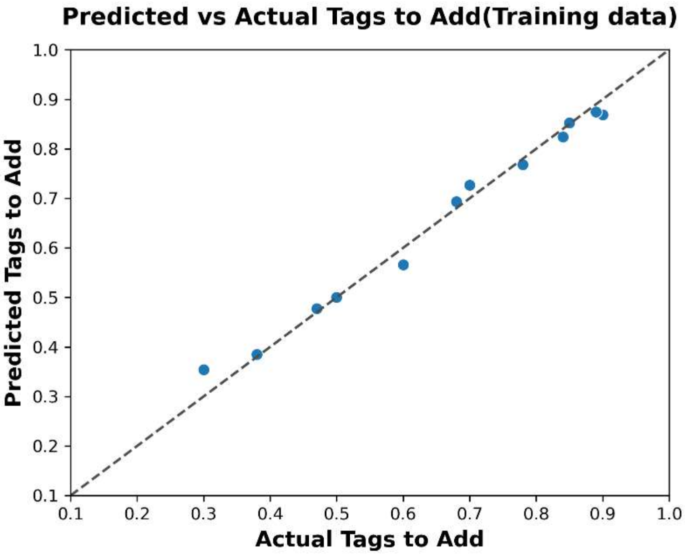

Figure 3. Random Forest Regression tags to Add training

Figure 3 demonstrates the performance of the random forest regression model in estimating "tags to add" using the training dataset. In this scatter plot, the actual values are plotted on the x-axis and the predicted values on the y-axis, with the diagonal dashed line representing a perfect prediction scenario. Most of the data points are tightly clustered around this line, revealing that the model accurately captures the training data patterns. The minimum scatter of points from the diagonal indicates low prediction error, high accuracy, and little or no signs of overfitting. This indicates that the model has successfully learned the relationships between key features such as image quality score and description length and their impact on the number of suggested tags. The random forest model proves effective in modeling complex, nonlinear relationships, making it a robust approach for this type of prediction task. While these results are promising, it is equally important to validate the model's generalization ability on new, unseen data. This feature is further explored in the following figure to ensure the reliability of the model in real-world applications.

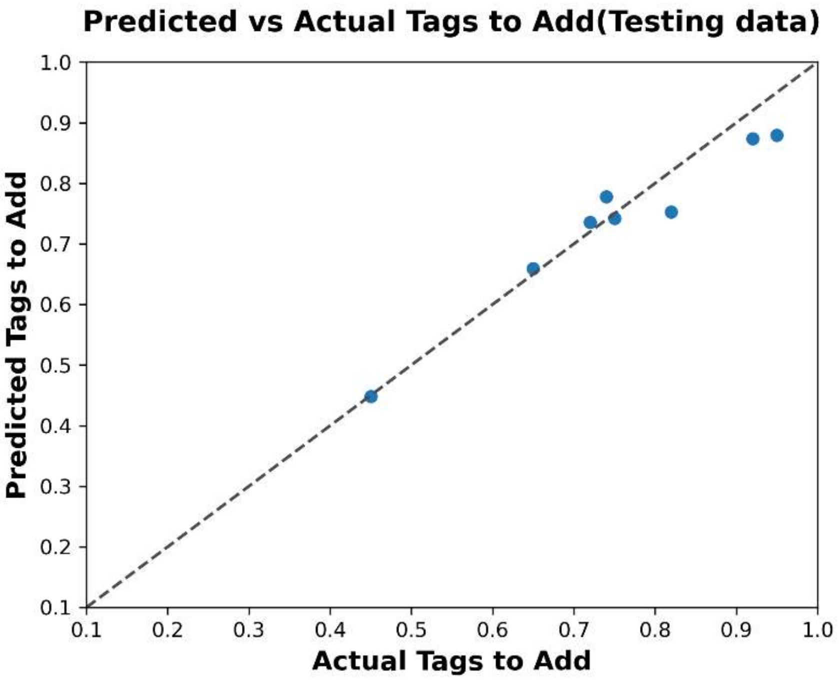

Figure 4. Random Forest Regression Tags To Add testing

Figure 4 illustrates how well the Random Forest Regression model performs on the test dataset, providing insight into its ability to generalize beyond the training data. Similar to Figure 3, this scatter plot compares the actual values of “tags to add” (x-axis) to the predicted values (y-axis), with the dashed diagonal line indicating the best predictions. Most of the points align closely with this line, indicating that the model maintains strong predictive accuracy even when applied to new, unseen data. Despite small deviations—especially at high tag values—the overall pattern indicates that the model generalizes well and maintains low prediction variance. This robust performance on the test data confirms that the model is not overfitting and is useful in optimizing features such as image quality and description length to predict tag recommendations. The alignment between training and testing results highlights the stability and performance of the model, reinforcing its potential for use in real-world situations that require consistent and accurate metadata tagging across diverse content inputs.

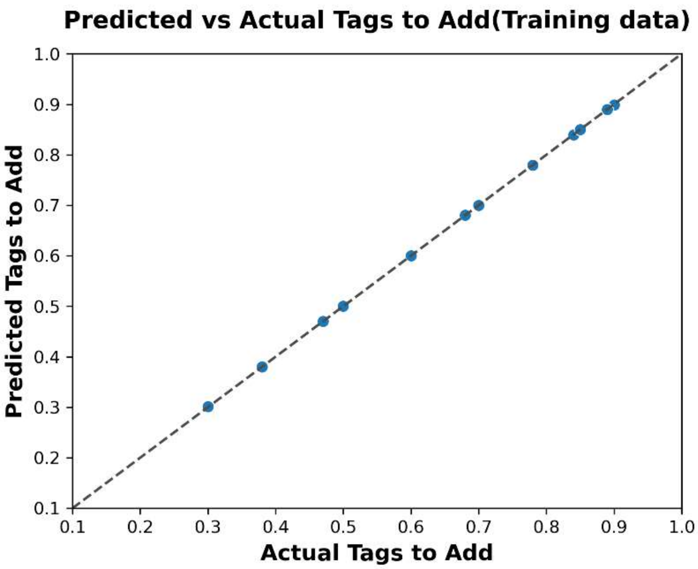

Figure 5. Xgboost Regression tags To Add training

Figure 5 shows how well the XGBoost regression model performs on the training data by comparing the actual values of the "tags to add" with their predicted counterparts. Most of the points closely follow the diagonal line, indicating perfect prediction accuracy. This tight clustering demonstrates the model's ability to accurately learn patterns and relationships within the training data, especially those involving key features such as description length, image quality score, and existing tag confidence. The results indicate a well-calibrated model with minimal prediction error and low bias. The nearly perfect fit underscores XGBoost's ability to handle complex data structures. However, exceptionally close alignment also requires caution, as such accuracy may indicate overfitting. Therefore, it is crucial to evaluate its generalization to unseen data.

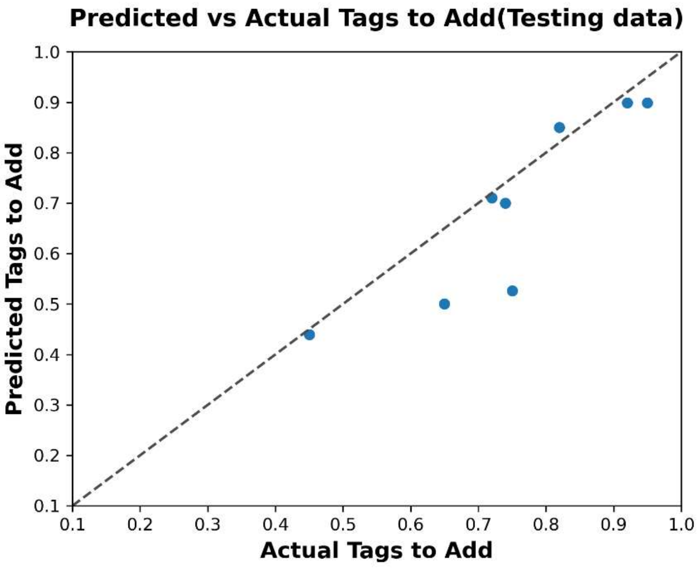

Figure 6. Xgboost Regression tags To Add testing

Figure 6 illustrates how the XGBoost model performs on the test data, providing insight into its generalization ability. Although the predictions mostly follow the diagonal line, especially for mid-range values, there are small deviations where underestimation occurs. Despite this, the overall trend is consistent with the actual values, indicating that the model maintains strong predictive accuracy. This consistency between predicted and actual results confirms the robustness and suitability of the model for applications that require reliable tagging. Although its performance on the unseen data is slightly lower than on the training set, it still demonstrates good generalization, supporting its performance in real-world situations.

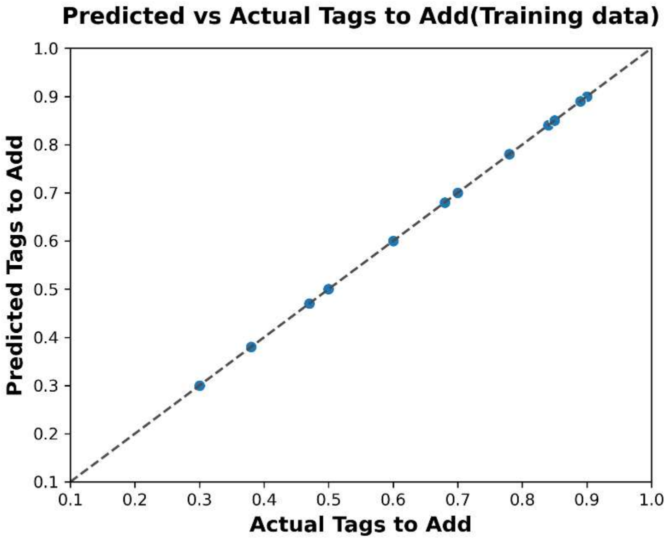

Figure 7. Decision Tree Regression tags to Add training

Figure 7 depicts the results of the decision tree regression model on the training dataset. The data points lie almost exactly on the diagonal line, indicating highly accurate predictions. Such performance is characteristic of decision trees, which closely fit the training data by partitioning it into specific partitions. While this can lead to better accuracy during training, it often signals overfitting, where the model memorizes the data rather than learning common patterns. This raises concerns about its ability to perform well on new or unseen data. While the results suggest that the model has captured all the relationships in the training set, its true performance can only be assessed by testing it on separate data sets

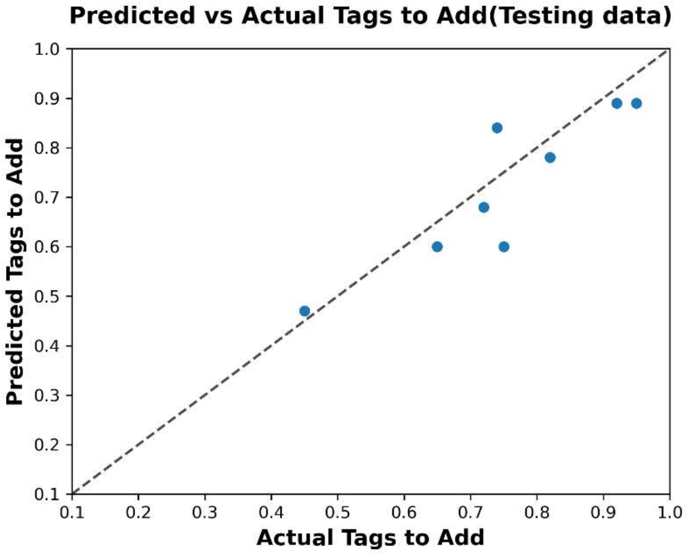

Figure 8. Decision Tree Regression tags To Add testing

Figure 8 illustrates the performance of the decision tree regression model on the test dataset. The scatter plot shows the actual "tags to add" values against the model's predictions, which show significant deviations from the best-fit diagonal reference line. These scattered points highlight the reduction in accuracy compared to the training results. Although the general direction of the data is consistent with the expected trend, the increased scatter indicates that the model struggles to generalize to the unseen data. This performance drop is characteristic of overfitting, especially common in unpruned decision trees. While the model performs well on training data by memorizing patterns, its adaptability is limited when applied to new inputs, indicating the need for pruning or ensemble techniques to improve its predictive reliability in real-world situations.

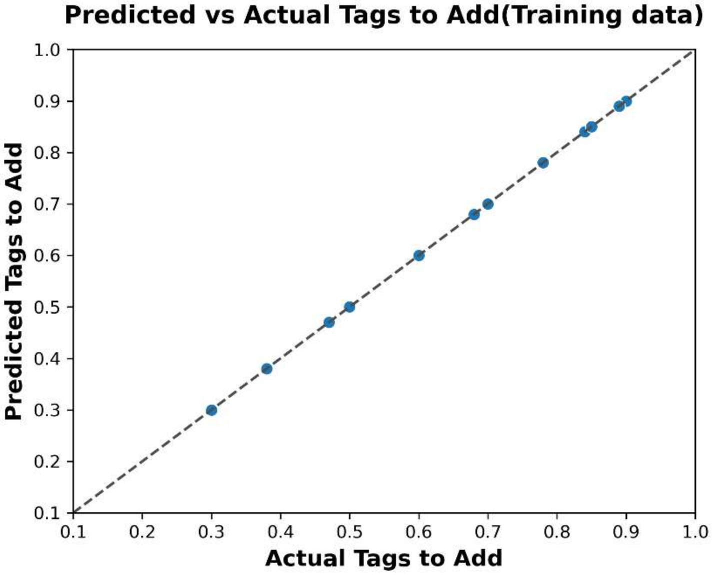

Figure 9. Gradient Boosting Regression tags to Add training

Figure 9 shows the predictions of the Gradient Boosting Regression model on the training dataset. The close clustering of points on the diagonal indicates exceptional accuracy and reflects the model’s strength in learning complex relationships between features. As a continuous ensemble method, Gradient Boosting iteratively improves its predictions by correcting for previous errors, resulting in a finely tuned fit. The near-perfect alignment in this graph demonstrates that the model effectively captures the structure of the training data. However, this high level of accuracy also calls for caution, as it can indicate overfitting, which reinforces the importance of evaluating performance on separate test data to ensure generalization.

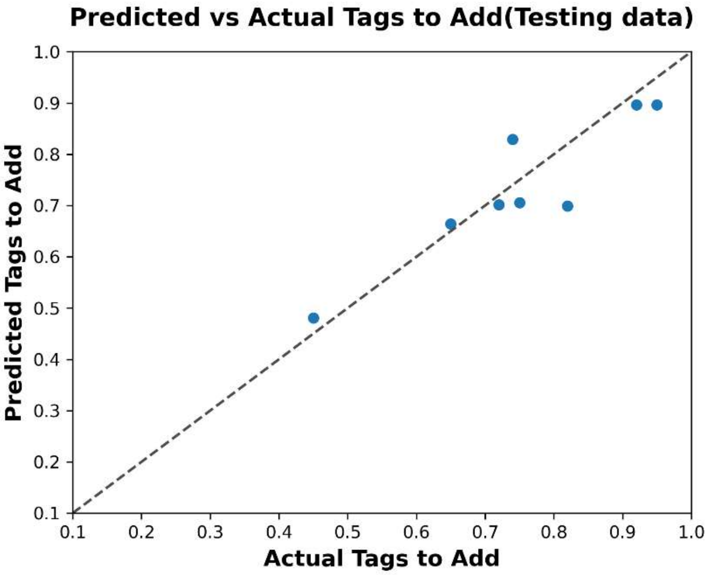

Figure 10. Gradient Boosting Regression tags to Add testing

Figure 10 evaluates how well the Gradient Boosting model generalizes when applied to test data. The scatterplot reveals that most predictions closely follow the diagonal line, indicating a strong fit between the actual and predicted values. Despite some minor discrepancies – especially in the mid-range – the model maintains an overall consistent pattern, indicating good generalization capabilities. This indicates that the model has effectively balanced learning and regularization, avoiding significant overfitting. The relatively low prediction error and constant variance confirm that gradient boosting remains a reliable approach for handling complex regression tasks such as automatic metadata tag suggestion.

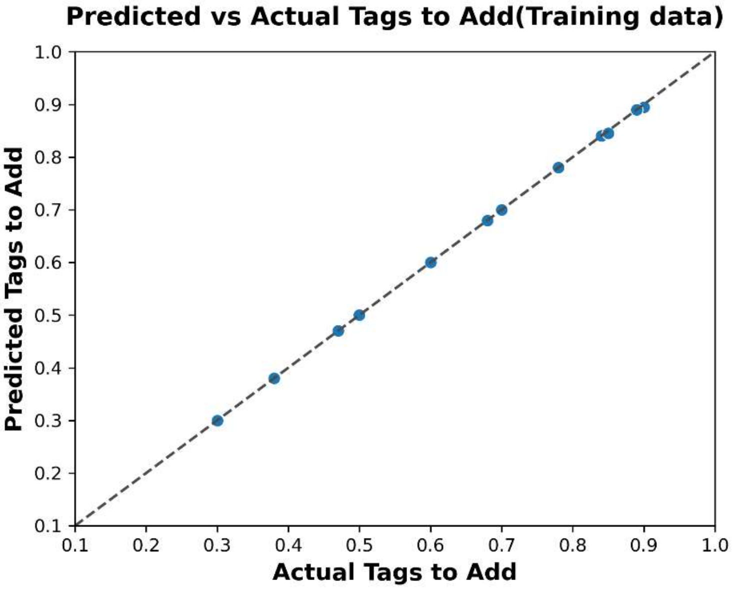

Figure 11. Adaboost Regression tags To Add training

Citation: PK Kanumarlapudi, “AI-Powered Product Metadata Enrichment through a Hybrid Approach Combining Semantic Web and Machine Learning” Journal of Business Intelligence and Data Analytics., 2025, vol. 2, no. 2, pp. 1–17. doi: https://dx.doi.org/10.55124/jbid.v2i2.250

Figure 11 illustrates a scatter plot showing the relationship between the actual and predicted values for “tags to be added” using the adaboost regression model on the training dataset. The points are tightly aligned with the 45-degree dashed line, which represents a nearly perfect relationship between the predicted and actual outcomes. This close alignment indicates that the model has effectively captured the underlying trends of the training data, achieving high accuracy with minimal error. the compact distribution of points on the reference line further emphasizes the strong fit of the model. however, such nearly perfect accuracy on the training data can also be a sign of overfitting, where the model performs exceptionally well on known data but may struggle to maintain the same performance on new, unseen inputs.

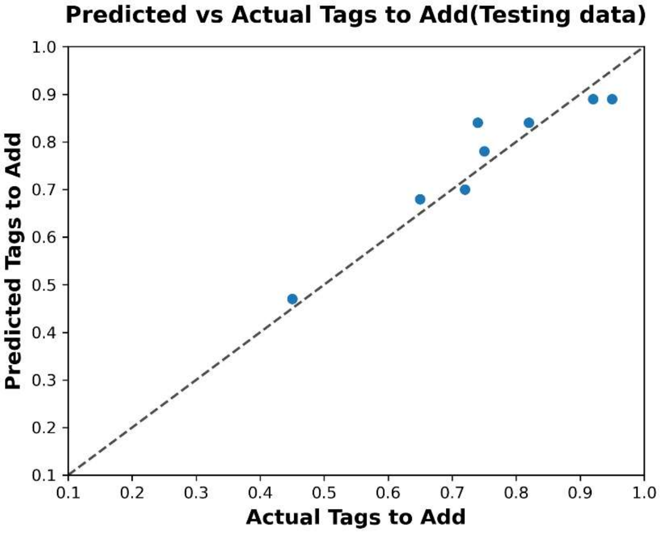

Figure 12. Adaboost Regression tags To Add testing

Figure 12 depicts how the AdaBoost model performs on the test dataset. Although the predicted values generally align with the 45-degree reference line, there is a bit more variability compared to the training plot. Some predictions fall above or below the line, indicating moderate discrepancies between the actual and predicted values. Despite these deviations, the model still demonstrates good predictive ability, indicating that it generalizes reasonably well to unobserved data. The slight scatter of the points indicates a small amountofoverfitting.

Conclusion

A study of metadata tagging enrichment in the e-commerce domain highlights the powerful synergy between semantic web technologies and machine learning methods. This research illustrates how integrating structured data standards with advanced regression models can significantly improve the efficiency, scalability, and accuracy of product tagging. Through the evaluation of five regression algorithms: Random Forest, XGBoost, Adaboost, Gradient Boosting, and Decision Tree, the study demonstrates that machine learning can effectively automate and refine metadata enrichment processes. Among the models tested, Random Forest Regression emerged as the most stable and generalize able, achieving high accuracy while avoiding the pitfalls of overfitting. Gradient Boosting and Adaboost produced robust results, while XGBoost and Decision Tree models, especially when unpruned, showed a tendency toward overfitting, reducing their reliability on new, unseen data. These results emphasize the need for proper regularization, cross-validation, and model optimization when building robust metadata enrichment systems. The results revealed a clear positive relationship between the quality of metadata enrichment and the richness of visual and textual content. Products with high-quality images and detailed descriptions consistently had better tag prediction results, reinforcing the value of detailed input data in AI-driven tagging. Furthermore, the study underscores the role of semantic frameworks such as schema.org and JSON-LD, along with ontology-based integration, in improving metadata quality. By aligning product data with standardized vocabularies and connecting to external knowledge sources, metadata becomes more meaningful, machine-readable, and interoperable. This semantic structure not only supports better search accuracy and personalized user experiences, but also facilitates broader data integration and reuse across platforms, especially in open data and digital commerce ecosystems.

References:

1. Ristoski, Petar, and Peter Mika. "Enriching product ads with metadata from html annotations." In The Semantic Web. Latest Advances and New Domains: 13th International Conference, ESWC 2016, Heraklion, Crete, Greece, May 29--June 2, 2016, Proceedings 13, pp. 151-167. Springer International Publishing, 2016.
2. de Castro, Bruno Portes Costa, Henrique Fernandes Rodrigues, Giseli Rabello Lopes, and Maria Luiza Machado Campos. "Semantic enrichment and exploration of open dataset tags." In Proceedings of the 25th Brazillian Symposium on Multimedia and the Web, pp. 417-424. 2019.
3. Costa, Ruben, Paulo Figueiras, Ricardo Jardim-Gonçalves, José Ramos-Filho, and Celson Lima. "Semantic enrichment of product data supported by machine learning techniques." In 2017 international conference on engineering, technology and innovation (ice/itmc), pp. 1472-1479. IEEE, 2017.
4. Ristoski, Petar, Petar Petrovski, Peter Mika, and Heiko Paulheim. "A machine learning approach for product matching and categorization: Use case: Enriching product ads with semantic structured data." Semantic web 9, no. 5 (2018): 707-728.
5. Alemu, Getaneh. "A Theory of Metadata Enriching and Filtering." Libri 66, no. 4 (2016): 251-262.
6. Silvello, Gianmaria, Georgeta Bordea, Nicola Ferro, Paul Buitelaar, and Toine Bogers. "Semantic representation and enrichment of information retrieval experimental data." International journal on digital libraries 18 (2017): 145-172.
7. De Nart, Dario, Carlo Tasso, and Dante Degl'Innocenti. "Users as Crawlers: Exploiting Metadata Embedded in Web Pages for User Profiling." In UMAP Workshops. 2014.
8. Tuarob, Suppawong, Line C. Pouchard, and C. Lee Giles. "Automatic tag recommendation for metadata annotation using probabilistic topic modeling." In Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries, pp. 239-248. 2013.
9. Pellegrini, Tassilo. "Semantic metadata in the publishing industry–technological achievements and economic implications." Electronic Markets 27, no. 1 (2017): 9-20.
10. Lin, Shao-En, Brian Liu, Miao-Chen Chiang, Ming-Yi Hong, Yu-Shiang Huang, Chuan-Ju Wang, and Che Lin. "BETag: Behavior-enhanced Item Tagging with Finetuned Large Language Models." In Proceedings of the ACM on Web Conference 2025, pp. 4996-5009. 2025.
11. Wang, Geng, Zhiqiang Lyu, and Xiaoyu Li. "An optimized random forest regression model for li-ion battery prognostics and health management." Batteries 9, no. 6 (2023): 332.
12. Afzal, Asif, Abdul Aabid, Ambareen Khan, Sher Afghan Khan, Upendra Rajak, Tikendra Nath Verma, and Rahul Kumar. "Response surface analysis, clustering, and random forest regression of pressure in suddenly expanded high-speed aerodynamic flows." Aerospace Science and Technology 107 (2020): 106318.
13. Du, Xishihui, Zhaoguo Wang, and Yan Wang. "The spatial mechanism and predication of rural tourism development in China: a random forest regression analysis." ISPRS international journal of geo-information 12, no. 8 (2023): 321.
14. Abdel-Rahman, Elfatih M., Fethi B. Ahmed, and Riyad Ismail. "Random forest regression and spectral band selection for estimating sugarcane leaf nitrogen concentration using EO-1 Hyperion hyperspectral data." International Journal of Remote Sensing 34, no. 2 (2013): 712-728.
15. Mittapally R (2023). Evaluating MicroStrategy Mobile and Competing Business Intelligence Solutions: A Multi-Criteria Decision-Making Approach. J Comp Sci Appl Inform Technol. 8(1): 1-9.
16. Yang, Yingbao, Chen Cao, Xin Pan, Xiaolong Li, and Xi Zhu. "Downscaling land surface temperature in an arid area by using multiple remote sensing indices with random forest regression." Remote Sensing 9, no. 8 (2017): 789.
17. Kalayci, Hakan, Umut Engin Ayten, and Peyman Mahouti. "Ensemble□based surrogate modeling of microwave antennas using XGBoost algorithm." International Journal of Numerical Modelling: Electronic Networks, Devices and Fields 35, no. 2 (2022): e2950.
18. Kiangala, Sonia Kahiomba, and Zenghui Wang. "An effective adaptive customization framework for small

manufacturing plants using extreme gradient boosting-XGBoost and random forest ensemble learning algorithms in an Industry 4.0 environment." Machine Learning with Applications 4 (2021): 100024.

19. Wang, Guangchao, Kun Liu, Hui Chen, Yusheng Wang, and Qingtian Zhao. "A high-precision method of flight arrival time estimation based on xgboost." In 2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT, pp. 883-888. IEEE, 2020.
20. Mittapally R (2024). Intelligent Framework Selection: Leveraging MCDM in Web Technology Decisions. J Comp Sci Appl Inform Technol. 9(1): 1-9.
21. Zou, Miao, Wu-Gui Jiang, Qing-Hua Qin, Yu-Cheng Liu, and Mao-Lin Li. "Optimized XGBoost model with small dataset for predicting relative density of Ti-6Al-4V parts manufactured by selective laser melting." Materials 15, no. 15 (2022): 5298.
22. Inoue, Tomoo, Daisuke Ichikawa, Taro Ueno, Maxwell Cheong, Takashi Inoue, William D. Whetstone, Toshiki Endo, KuniyasuNizuma, and Teiji Tominaga. "XGBoost, a machine learning method, predicts neurological recovery in patients with cervical spinal cord injury." Neurotrauma reports 1, no. 1 (2020): 8-16.
23. Rakhra, Manik, Priyansh Soniya, Dishant Tanwar, Piyush Singh, Dorothy Bordoloi, Prerit Agarwal, Sakshi Takkar, Kapil Jairath, and Neha Verma. "WITHDRAWN: Crop price prediction using random forest and decision tree regression: -A review." (2021).
24. Bensic, Mirta, Natasa Sarlija, and Marijana Zekic□Susac. "Modelling small□business credit scoring by using logistic regression, neural networks and decision trees." Intelligent Systems in Accounting, Finance & Management: International Journal 13, no. 3 (2005): 133-150.
25. Fonarow, Gregg C., Kirkwood F. Adams, William T. Abraham, Clyde W. Yancy, W. John Boscardin, and ADHERE Scientific Advisory Committee. "Risk stratification for in-hospital mortality in acutely decompensated heart failure: classification and regression tree analysis." Jama 293, no. 5 (2005): 572-580.
26. Chen, Wei, and Zifan Yang. "Landslide susceptibility modeling using bivariate statistical-based logistic regression, naïve Bayes, and alternating decision tree models." Bulletin of Engineering Geology and the Environment 82, no. 5 (2023): 190.
27. Hong, Wandong, Lemei Dong, Qingke Huang, Wenzhi Wu, Jiansheng Wu, and Yumin Wang. "Prediction of severe acute pancreatitis using classification and regression tree analysis." Digestive diseases and sciences 56 (2011): 3664-3671.
28. Mittapally R (2023). Evaluating Business Intelligence Alternatives: COPRAS vs Traditional Models in MicroStrategy. J Comp Sci Appl Inform Technol. 8(1): 1-9.
29. Rizkallah, Lydia Wahid. "Enhancing the performance of gradient boosting trees on regression problems." Journal of Big Data 12, no. 1 (2025): 35.