International Journal of Recent Advances in Multidisciplinary Topics
Volume 5, Issue 5, May 2024
https://www.ijramt.com | ISSN (Online): 2582-7839

Recommender System for Website Content Management System Selection

Dmitrij Kolodynskij*

Faculty of Fundamental Sciences, Vilnius Gediminas Technical University, Vilnius, Lithuania

***Abstract*: The aim of this study is to automate the process of selecting content management systems based only on company profile data. In this study, a web-browser scenario has been developed to extract enterprise data from publicly available sources. The study was carried out in the Republic of Lithuania. The analysis of the data mining models showed that by applying the Decision Tree algorithm to the classification problem, it is possible to achieve model accuracy of more than 90% in predicting the CMS used by companies based on company profile data.**

***Keywords*: content management system, data mining, machine learning, recommender system, web scraper.**

1. Introduction

Over the last two decades, the concept of CMS has evolved due to the high demand for online data management and support [1], [2]. This is why it is not only important to create a website, but also to consistently manage the displayed information. CMS allows to extend the content of a web-based system by applying user-friendly tools, to create and manage the rights and roles of the users of the system, to use plug-ins to enhance the system and to add functionality, etc [2]. CMS has a strong online presence, with around 70 million websites worldwide built using its products [3]. According to statistics on websites using CMS [4], WordPress is the most popular technology in the CMS category, with about 63% of websites based on the system. 6% of online projects are developed using Shopify CMS. Wix, Squarespace, Joomla and Drupal are used in about 3% of the websites respectively, while 20% of the websites are realised with other CMS. Wix, Squarespace, Joomla and Drupal are used in about 3% of the websites respectively, while 20% of the websites are realised with other CMS.

There are many commercial and free CMS products available on the market, which vary in terms of features – level of complexity of use, purpose, advantages, disadvantages etc. The use of a particular CMS product depends on the user's needs and the planned web project. As information technology is constantly evolving, the quantity and quality of CMS products is constantly changing. It is therefore important to identify trends in the use of CMS in companies, depending on the field of activity, to analyse the data and to develop a framework for selecting the most appropriate CMS for the company profile.

This study has analysed similar studies to examine the data analysis methods used and to define the company profile data needed for the study. A search of company data sources and an analysis of open information was carried out to identify the most appropriate data source. Further, the steps of data crawling from a publicly available source were defined, a data crawling algorithm was developed and the analysis of the resulting data was carried out in the form of data preparation and modelling. An automatic classification model builder was applied and manual k-NN, Decision Tree and Naive Bayes models were built. The results of the study were evaluated to identify the classifier that most accurately identifies the CMS from the company profile data.

2. Related Works

In order to properly select CMS data mining methods and to assess the contribution of other researchers, it is important to carry out an analysis of research in the field of CMS data mining. The literature review examined research papers that applied different data analysis methods.

Skatikiene et al (2017) [5] – the study analyses the most popular CMSs, presents CMS awareness survey data from different categories of Lithuanian websites, the results of the CMS survey, and identifies criteria and recommendations for the selection of CMSs for websites. The results of the study:

- A website built on a modern and popular CMS has the advantage of new modules, user-friendly content management and a large community.
- If the user has no programming knowledge, WordPress CMS is recommended for a corporate website.
- The most popular e-commerce CMS in the Lithuanian market – WordPress, PrestaShop, OpenCart.
- For more complex platforms, Drupal CMS is recommended for development and management.

Hwang et al (2016) [6] – the study examines how consumers' proactive motivation levels towards information use and their attitudes towards newly introduced technologies in an organisational setting, either before the technology is launched or before the implementation phase, affect the adoption beliefs of CMS. This study argues that information proactivity influences adoption beliefs such as perceived ease of use and

*Corresponding author: d.kolodynskij@gmail.com

perceived usefulness. The results show that information activity is a significant determinant of system users' perceived ease of use, but not of perceived usefulness prior to deployment.

Laumer et al (2017) [7] – the paper explores a critical aspect of how organisations work - the use of information provided by information systems such as corporate CMS in work tasks. Based on a model of information systems success, interviews and an empirical study conducted among corporate CMS users in a financial services provider company, the authors identify two dimensions of information quality: context and representation. Furthermore, the study reveals that these dimensions, together with the quality of the system, are important in determining end-user satisfaction, which influences the emergence of circumvention strategies. This study provides recommendations for organisations to implement the most appropriate countermeasures based on the importance of the quality of the context or representation information.

Drivas et al (2021) [8] – the authors analysed the performance of 341 websites based on three different factors: content creation, speed and security. The first phase presented a statistically robust and consistent scoring scheme to evaluate the SEO performance of library, archive and museum websites, integrating more than 30 variables. The second phase involved a descriptive data synthesis for the initial evaluations of the websites' performance in each factor. In the third stage, predictive regression models were developed to understand and compare the SEO performance of three different CMSs - Drupal, WordPress and adapted approaches implemented by the LAM sites.

The analysis of the studies shows that the articles use different research methods - regression, statistical analysis, etc. There is also a lack of research on the use of CMS by companies. Skatikienė et al (2017) [5] analysed CMS and their use in companies, but the data sample used was only 100 companies, so a small part of the Lithuanian market was covered. Moreover, the proposed recommendations for CMS were based on analysis of CMS, usage statistics and surveys, but not on company profile data.

In order to train a model to identify the CMS of companies, some company profile data is needed to reveal the needs, resources and capabilities of companies. The analysis of the studies has led to the definition of data that can be useful for this purpose:

- *Company size.* The size of a company, measured in terms of number of employees, revenue or other indicators, can help determine which CMSs are appropriate for a particular company. Large companies may require more complex and powerful systems, while small companies may be suited to simpler and cheaper alternatives.
- *Field of activity.* The scope of a company's activities can affect its CMS needs. For example, in an online shop, security of payment functionality, support for different payment systems, product management, etc. are important. Meanwhile, a news portal will have content management capabilities, user role mechanisms, multilingualism, etc.
- *Geographical location.* The geographical location of companies can affect the requirements of the CMS, such as support for different languages, compliance with local legislation, infrastructure requirements.
- *Staff competence.* The choice of the right CMS depends on the level of skills and competences of the staff. Companies that have staff with technical knowledge and experience have the ability to implement more complex and flexible systems, compared to companies with only basic (entry-level) skills. In such cases, the CMS must have a user-friendly interface and not require specific programming knowledge.
- *Technological infrastructure.* The technological infrastructure of a company can influence which CMS will be compatible with the demands placed on it. This may include the devices used, software, internet connection speeds, server service providers.
- *The budget.* The financial resources allocated to the company's IT department determine the choice of CMS. Companies with large resources have more options, with more expensive and powerful solutions available, while smaller companies are interested in more cost-effective alternatives.
- *Integration requirements.* The technologies and solutions used in the company may require infrastructure to be aligned with the planned CMS. This may include business and resource management systems, email systems or other software integrations.
- *Access control and security.* Each company has individual security requirements and user authorisation. Depending on the needs of the company and its internal procedures, data encryption, user authentication and authorisation, two-factor authentication and other security features may be required.

3. The Proposed Methodology of the Recommender System

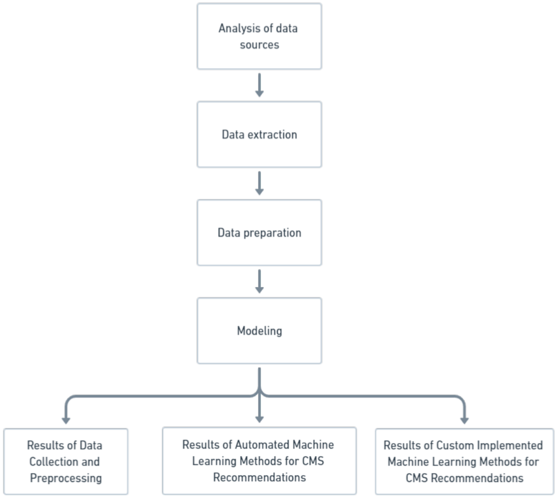

Analysis of data sources
Data extraction
Data preparation
Modeling
Results of Data Collection and Preprocessing
Results of Automated Machine Learning Methods for CMS Recommendations
Results of Custom Implemented Machine Learning Methods for CMS Recommendations

Fig. 1. Stages of the study

With regard to the stages of this study (Fig. 1), firstly, the work involved a search for the source of data on the company. Then, a data extraction process was developed using the previously selected source. The resulting data was then prepared for modelling. After the modelling process, the results of the modelling were reviewed.

A. Extraction of Dataset

A search of data sources and analysis of open information on enterprises in the Republic of Lithuania was carried out. Four information portals publishing a directory of companies and providing company search and data access were selected:

- Sodra.lt open data [9]. The system publishes indicators in JSON and CSV formats by year and month. Available data that could be useful for the training of the model include average wages, number of insured persons, amount of national social security contributions, insurance premium debt, insurance premium deferred debt. Advantages – no programming of the web-browser is required as a list of all enterprises is provided. Disadvantages: insufficient data for the analysis. Information on companies' websites is not provided
- Okredo.lt web portal [10]. The portal publishes a list of companies' activities and provides a search facility for companies. The name, address, company code, field of activity, status, date of registration and number of employees of the company are provided. The openly available data is not sufficient to train the model.
- Rekvizitai.lt web portal [11]. The portal publishes companies' activities and provides a search engine for companies. The company information pages provide a wealth of information and useful data: company website, number of employees, average salary, median salary, sales revenue, net profit, number of brands, share capital, age of company. The portal uses a security code, one of the objectives of which is to protect against web browsers and to reduce server loads due to the high volume of requests. The security code complicates and significantly slows down the entire data extraction process.
- Info.lt portal [12]. The portal publishes companies' activities, provides a search engine for companies and provides useful information about the company - website, date of registration, number of employees and average salary. An analysis of the security measures used showed that the system does not use a security code, but if a certain number of requests are made in a short period of time, the client's IP address is blocked for a limited period.

After assessing the availability of information on web portals, the amount of data available and the safeguards in place, the data source most suitable for the task at hand was selected. The model was developed using company profile data and the CMS used on the websites. The work addressed a classification task, so the dataset was divided into classes according to different types of CMS. Using the extracted dataset, a model was built and trained with the aim of recommending the most suitable CMS given the company profile data. It is important to note that not all of the defined data is available on the internet, so it should be taken into account what information is openly available on online portals. The extraction of the dataset was divided into two phases: extraction of the company profile data from the web portal and identification of the CMS used by the companies.

Web Scraper – software or script used to automatically collect and extract content from websites. Web scrapers allow fast access to large amounts of information and useful data on the Internet [13]. The Web Scraper script may face certain challenges during its operation, such as changes in the structure of a web page, security measures such as captcha or robots.txt files, ethical and legal issues related to the extraction of unauthorised data. For these reasons, it is important to take into account potential risk factors when searching for and analysing the source of the data and when designing the algorithm of the web browser.

The extraction of the company profile dataset consisted of several steps: analysis of the data sources; construction of the extraction steps; implementation of the web browser script; and running the script. Taking into account the structure of the web portal, three phases of the web browser were developed:

- *Extraction of the fields of activity of enterprises*, i.e., scanning of all the fields of activity of enterprises, with links to the list of enterprises belonging to a specific field. The aim was to carry out the subsequent extraction of links to the company pages in a consistent way, according to the categories of companies. The data were stored in a table in the database of the fields of activity.
- *Extracting company names and links*. Using scanned links to company listings, company names and links to web portal pages were extracted. The data was stored in a table in the enterprise database.
- *Extraction of company profile data*. Using links to web portal pages with company information, detection and extraction of company profile data was performed.

Based on the stages of the web browser, three different processes were created and launched individually. Dividing the data extraction process into stages made it possible to manage and control the execution of the processes, identify and solve problems in time, correct errors and ensure high quality of results.

Web browsing is a complex process that requires planned strategies and techniques to ensure data accuracy and reliability. The following steps are important in this process:

- *Changing User-Agent values.* The User-Agent header of the HTTP request defines information about the user's software and operating system that is retrieved by the network server. A changing User-Agent header is used to reduce the chance of automated content extraction script detection and to simulate human browsing. In this way, the script is protected from potential IP address blocking or other server security measures.

- *IP Rotation.* Most websites use various security measures to detect and block automated content extraction bots and scripts. One of the browser's ways of restricting the browser is to block the IP address. In order to avoid blocking of IP address, measures of changing IP address are used. Various technologies are used to achieve this goal, such as proxies or VPN.
- *Error and exception handling.* Errors and exceptions may occur during the data extraction process, such as changes in web page structure, server errors, or Internet connection problems. Anticipating and managing errors and exceptions ensures fault tolerance and uninterrupted execution of the web browser process.
- *Use of random pauses.* The use of random pauses between requests during script process execution is intended to simulate human browsing and reduce the likelihood of automated data mining being detected on the source server. Additionally, a web browser script with random pauses significantly reduces IP address blocking and source server loads.
- *Data processing and cleaning.* The resulting data may be inaccurate or contain redundant information. During data processing, data must be cleaned and transformed in such a way that they are suitable for storage and further data analysis.
- *Data storage.* Extracted data must be stored, ensuring further convenient use. Data can be stored in databases, CSV files, or other data structures that allow easy access and review of the data.

In order to identify the CMS used by companies, the third-party tool whatcms.org [14], which provides a free API service, was used. A script was created to set up the companies' CMS using the API tool and update the records in the database table

B. Data Analysis

The data analysis consisted of a series of steps to prepare the data and to identify the most effective methods for forecasting the CMS based on company data. The data analysis was divided into two phases - data preparation and building prediction models.

In the data preparation phase, the RapidMiner Studio v9.10 software [15] was used to analyse the dataset and define the techniques and methods that were used to prepare the dataset for modelling. In the modelling phase, sub-processes of the models were developed by applying different parameters of the data mining algorithms and evaluating the resulting models. Finally, the most appropriate model is presented based on the results of the simulations performed.

1) Data Preparation

When analysing the data and solving the classification task, certain attributes may be undesirable or have no value, for the following reasons:

- *High number of omitted values.* Attributes with a large number of omitted values may not be informative enough for model training. In such cases, the attribute is removed to reduce model noise and increase accuracy.
- *Low variation in values.* Attributes with low variance (for example, when most or all values are the same) reduce the accuracy of the model due to insufficient information to solve the classification task.
- *Unimportant attributes.* Attributes may be irrelevant in the context of a classification problem. For example, the name of a customer will not influence the prediction of product purchase. Removing irrelevant attributes allows you to focus on more relevant data and increase the accuracy of the model.
- *Confidentiality and privacy.* In some cases, the removal of attributes is necessary due to confidentiality or privacy requirements. For example, personal identification numbers, passwords or personal data should be removed from the analysis to ensure customer privacy and to comply with legal requirements.

During the data preparation stages, the values of the dataset were reviewed to detect and remove outliers, as outliers can affect the accuracy and overfitting of the model. Normalisation was applied to improve the accuracy and efficiency of data mining algorithms using distance measurements [16].

Cross-Validation, often referred to as k-fold cross-validation, is an effective statistical method used in machine learning and statistical modelling to assess the performance of predictive models. It is a resampling technique that provides a more accurate assessment of model performance based on the ability to generalise unseen data. Cross-validation reduces the risk of overlearning that occurs when a model is overcomplex and effective with training data but inaccurate with new, unseen data. Using different combinations of training and testing data provides a more reliable assessment of model performance by ensuring that the model is evaluated on a wide range of data [16], [17].

2) Building prediction models

Auto Model is an extension to RapidMiner Studio software that allows automation of the process of building and verifying a data training model [18]. The Auto Model extension was used in the study to see if an automated model training tool would be suitable for the task at hand. Advantages of the extension:

- *Automatic data preparation.* Auto Model algorithms aim to automatically perform a wide range of data preparation tasks, such as missing value detection, data normalisation, etc.
- *Model selection and training.* The extension automatically selects the most appropriate machine learning algorithms based on the dataset and the problem to be solved. As hyperparameter identification can be a complex and time-consuming process, Auto Model aims to automatically perform the identification process.
- *Model evaluation.* The Auto Model process results in model evaluation reports that help you understand model performance and plan improvement actions.

- *Time costs.* Auto Model automates tasks that require time-consuming and technical expertise.

Auto Model is a useful tool, but should not be used as the only way to create machine learning models. The problem must take into account the characteristics of the dataset and the possibilities of adapting the models to specific problems. Although the Auto Model extension aims to automate and speed up the model building process, it has limitations:

- *Scalability of the model.* Auto Model's capabilities may be limited when dealing with complex tasks or exceptional datasets that require advanced data engineering processes. This can make it difficult to achieve optimal results in more complex tasks.
- *Limitations of model optimisation.* Although the Auto Model process automatically sets hyperparameters during execution, more complex aspects of optimisation may not be available. This may limit the scope for model improvement.

K-Nearest Neighbours (K-NN). The classifier is based on the idea that similar objects are closer to each other in space. The performance of the k-NN algorithm depends on several important factors, such as the choice of an appropriate k parameter, the choice of a distance measure such as Euclidean or Manhattan, as well as the adjustment of the data scale and dimension. In the modern scientific literature, there are many methods to optimise these factors, as well as modifications and 20 hybrid k-NN algorithms to improve the performance of the algorithm [19], [20]

The K-Nearest Neighbours algorithm depends on a pre-specified parameter K. A loop was designed to select the value of k with which the accuracy of the model would be highest. The iterators of the loop were set from 1 to 20.

The Decision Tree Method is a hierarchical structure made up of nodes and branches. The main component of a Decision Tree is the node, which can be root, internal or leaf. The internal nodes perform the evaluation of the feature criteria of the dataset, while the leaf nodes specify the final class.

During the training process of the Decision Tree classifier, the dataset is split into smaller subsets based on feature criteria. The partitioning continues until leaf nodes are reached in which all elements belong to the same class, or the node becomes too small for further partitioning. Splitting criteria [16]:

- *Gain ratio*. A modification of the information gain to target trials with multiple outcomes. This means that attributes with multiple values are preferred. The information gain is adjusted for internal partitioning information.
- *Information gain*. A metric that measures the expected entropy reduction due to feature sorting. Entropy is a measure that describes the inconsistency or uncertainty of a data set. The lower the entropy, the more information about the class can be obtained from the attribute. Used to build Decision Trees from the dataset in ID3 and C4.5 algorithms.
- *Gini index*. Another way to measure uncertainty or inconsistency in a dataset. Unlike entropy, the Gini index does not consider all possible classes and their probabilities, but instead focuses on the most common class. The Gini index is lower when the most frequent class makes up a larger proportion of the dataset, and higher when the classes are evenly distributed.

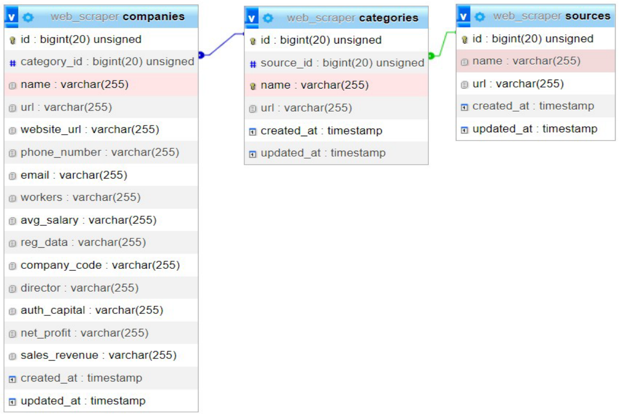
As the Decision Tree algorithm depends on a predefined depth parameter, a loop was designed to select the depth value with which the model accuracy would be highest. The iterators of the loop are set from 1 to 100. Three models were developed in the study with different partitioning settings – gain ratio, information gain and Gini index.

Naive Bayes. A method based on Bayes' theorem that incorporates the concept of conditional probability in classification. The advantages of the classifier are speed of computation, low computational resource requirements and efficiency with large datasets. However, due to the construction of the theorem, the algorithm does not perform well when a certain combination of values is missing in the training data. The Naive Bayes method is effective when the categories are simple, for example, for tasks where the keywords are features (e.g. spam detection), but the method does not work when the relationship between words is important (e.g. sentiment analysis) [16], [21].

4. Results and Discussion

A. Results of Data Collection and Preprocessing

The analysis of the data sources of the companies shows that the open data available in Sodra.lt and Okredo.lt is not sufficient for the analysis. The Rekvizitai.lt web portal publishes a lot of useful information, but security measures limit the possibilities of the web browser. The Info.lt portal provides less data compared to the Rekvizitai.lt web portal, but with a properly configured data extraction algorithm it is possible to access openly published company profile data and to carry out further analysis. The Info.lt web portal was found to be suitable for the task. Using the configured data extraction algorithm, openly published company profile data was accessed.



web_scraper **companies**
id : bigint(20) unsigned
category_id : bigint(20) unsigned
name : varchar(255)
url : varchar(255)
website_url : varchar(255)
phone_number : varchar(255)
email : varchar(255)
workers : varchar(255)
avg_salary : varchar(255)
reg_data : varchar(255)
company_code : varchar(255)
director : varchar(255)
auth_capital : varchar(255)
net_profit : varchar(255)
sales_revenue : varchar(255)
created_at : timestamp
updated_at : timestamp

web_scraper **categories**
id : bigint(20) unsigned
source_id : bigint(20) unsigned
name : varchar(255)
url : varchar(255)
created_at : timestamp
updated_at : timestamp

web_scraper **sources**
id : bigint(20) unsigned
name : varchar(255)
url : varchar(255)
created_at : timestamp
updated_at : timestamp

Fig. 2. MYSQL database schema for the web scraper

Three database tables (Fig. 2) – companies, categories and sources - were created to define the web browser algorithm. The

Table 1
Description of dataset attributes

Attribute name	Type	Missing values	Description
category	Qualitative, nominal	0	Name of the company's field of activity (e.g., Lawyers, Cafés)
name	Qualitative, nominal	0	Company name
url	Qualitative, nominal	0	Link to the portal's web page with company profile data
website_url	Qualitative, nominal	0	Link to corporate website
cms	Qualitative, nominal	2084	Name of content management system
phone_number	Quantitative, discrete	0	Company phone number
email	Qualitative, nominal	57	Company email address
workers	Quantitative, discrete	0	Number of employees in the company
avg_salary	Quantitative, continuous	0	Average salary in the company
reg_data	Qualitative, nominal	97	Date of company registration
company_code	Quantitative, discrete	0	Company code
director	Qualitative, nominal	1286	Name and surname of the company director

Sources table stored the names of the information parties, the categories table stored the company category data, and the companies table stored the company data.

The data analysis identified 6306 records in the dataset. The dataset includes 8 qualitative, nominal, 3 quantitative, discrete and one quantitative continuous attributes (Table 1). The attributes 'name', 'url', 'website_url', 'phone_number', 'email', 'reg_data', 'company_code' and 'director' are not relevant to the classification task and should therefore be removed from the dataset. The classification task was solved using the 'category', 'cms', 'workers' and 'avg_salary' attributes.

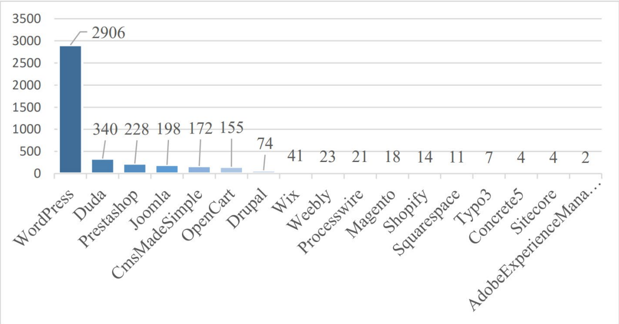


Fig. 3. Distribution of observations for the CMS attribute

During the data extraction phase, 17 different CMSs were identified that are used on company websites. Looking at the bar chart of the distribution of CMSs (Fig. 3), it can be seen that certain types of CMSs, such as Drupal, Wix, Weebly, Shopify and others, have a significantly lower frequency of occurrence compared to other types. Due to the low frequency of occurrence, the model will require more rules or classes to obtain higher accuracy and predict less frequent values, for example, in the case of the Decision Tree algorithm, a higher depth. Decision Tree rules constructed in this way, which generate nodes with two or three examples, will be of little practical use. Therefore, the aim of the prediction model building phase was to create a model with a small number of nodes corresponding to a large number of training samples. To achieve this, records with infrequently occurring values of the 'cms' attribute were combined into a new type Other.

Table 2
Attribute value ranges

Attribute name	Range of values
workers	[1; 1234]
avg_salary	[94,980; 6567,360]

Before the data cleaning process, the values of the dataset were reviewed to detect outliers (Table 2). It was observed that values of the ‘workers’ attribute greater than 500 are significantly out of line with the overall distribution of the data. Considering the distribution of values of the ‘avg_salary’ attribute, values greater than 6000 are outliers from the other observations. These outliers may affect the accuracy and overfitting of the model and were therefore removed during the data preparation stage. The ranges of values for the attributes 'workers' and 'avg_salary' were found to be significantly different, so normalization was applied to these attributes.

B. Results of Automated Machine Learning Methods for CMS Recommendations

According to the results of the Auto Model process (Table 3), the accuracy of all five machine learning algorithms used - Naive Bayes, Decision Tree, Deep Learning, Logistic Regression and Random Forest - are similar to each other, ranging between 68.3% and 69.0%. The Naive Bayesian model has the highest accuracy at 69.0%, while the Random Forest model has the lowest accuracy.

Table 3
Auto Model process results

Model	Accuracy	Training and testing time
Naïve Bayes	69.0 %	28 sec.
Decision Tree	68.9 %	24 sec.
Deep Learning	68.8 %	39 sec.
Logistic Regression	68.5 %	45 sec.
Random Forest	68.3 %	1 min. 21 sec.

The results showed that the dataset is complex and the defined task is not easily solvable using automated model training. The performance of different algorithms depends on the dataset and the problem to be solved, so in order to develop an efficient model, it is necessary to test several different algorithms by selecting appropriate parameters. As the Auto Model process did not identify a model that effectively solved the defined problem, further algorithms with different parameters were used to obtain a better result.

C. Results of Custom Implemented Machine Learning Methods for CMS Recommendations

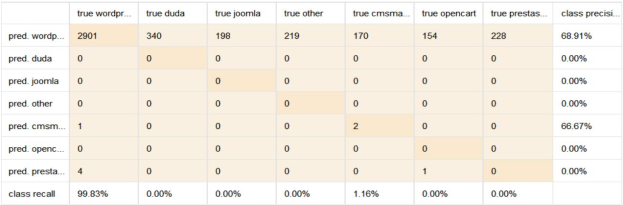
1) K-Nearest Neighbours model

Analysing the accuracy results of the K-Nearest Neighbours classification model (Table 4), it can be seen that the choice of the number of k has a small effect on the accuracy. The accuracy ranges from 68.5% with k = 10 to 68.9% with k = 17, 19 and 20. The lowest accuracy was obtained with k = 1 (51%). Considering the results obtained, it can be seen that the accuracy of the model reaches 69% and does not vary significantly with the number of k neighbours.

Table 4
K-Nearest Neighbours model accuracy by K

k	Accuracy
1	51.0 %
2	63.3 %
3	64.8 %
4	65.6 %
5	66.4 %
6	67.4 %
7	67.5 %
8	67.9 %
9	68.2 %
10	68.5 %
11	68.5 %
12	68.5 %
13	68.5 %
14	68.5 %
15	68.7 %
16	68.8 %
17	68.9 %
18	68.8 %
19	68.9 %
20	68.9 %

In order to test the ability of the model to predict different TVS values, a confusion matrix was developed (Fig. 4). The parameter k = 10 was chosen as the accuracy of the model does not change significantly as K increases. The results show that all classes of CMS systems, except WordPress, are null. The model predicts the WordPress type with 99.83% accuracy, but the other CMS cannot be identified or predicted with low accuracy.



	true wordpr...	true duda	true joomla	true other	true cmsma...	true opencart	true prestas...	class precisi...
pred. wordp...	2901	340	198	219	170	154	228	68.91%
pred. duda	0	0	0	0	0	0	0	0.00%
pred. joomla	0	0	0	0	0	0	0	0.00%
pred. other	0	0	0	0	0	0	0	0.00%
pred. cmsm...	1	0	0	0	2	0	0	66.67%
pred. openc...	0	0	0	0	0	0	0	0.00%
pred. presta...	4	0	0	0	0	1	0	0.00%
class recall	99.83%	0.00%	0.00%	0.00%	1.16%	0.00%	0.00%	

Fig. 4. K-Nearest Neighbours model confusion matrix

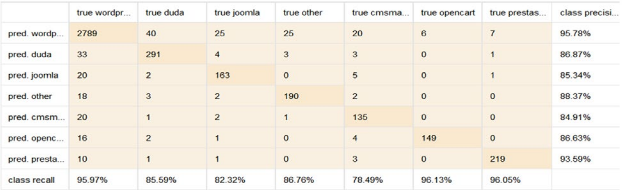
2) *Decision Tree model*

Table 5
Accuracy of Decision Tree models by parameter

Gain ratio		Information gain		Gini index	
Depth	**Accuracy**	**Depth**	**Accuracy**	**Depth**	**Accuracy**
100	81.6 %	69	93.6 %	42	93.9 %
98	81.5 %	66	93.6 %	80	93.9 %
99	81.3 %	34	93.6 %	83	93.8 %
90	81.3 %	74	93.5 %	69	93.8 %
95	81.3 %	99	93.5 %	49	93.7 %
94	81.3 %	52	93.5 %	90	93.7 %
92	81.3 %	80	93.5 %	73	93.7 %
96	81.3 %	21	93.5 %	37	93.7 %
89	81.2 %	63	93.5 %	75	93.7 %
91	81.2 %	37	93.5 %	66	93.7 %

Analysing the accuracy results of the Decision Tree classification model, sorted in descending order of accuracy (Table 5), it can be observed that the relative information gain decomposition criterion resulted in the lowest accuracy among all the models developed with different criteria. Moreover, the accuracy of the model is higher with increasing depth using the above criterion. Thus, in order to train a more accurate model, a large value of depth must be chosen, which increases the probability that the model will be overtrained or not practically applicable (due to the large number of rules).

Using the information gain and Gini index decomposition parameters, the models with the highest accuracies were obtained. Considering the depth to model accuracy ratio – the smallest depth with the highest model accuracy – the best fitting depth parameters were selected: information gain depth – 21, accuracy – 93.5%, Gini index depth – 37, accuracy 93.7%. Considering the confusion matrix (Fig. 5) of the model with the information gain partitioning parameter and a depth of 21, it can be seen that all the classes are predicted with high accuracies ranging from 96.13% to 78.49%.

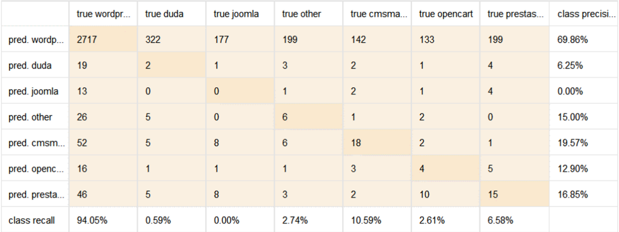


	true wordpr...	true duda	true joomla	true other	true cmsma...	true opencart	true prestas...	class precisi...
pred. wordp...	2789	40	25	25	20	6	7	95.78%
pred. duda	33	291	4	3	3	0	1	86.87%
pred. joomla	20	2	163	0	5	0	1	85.34%
pred. other	18	3	2	190	2	0	0	88.37%
pred. cmsm...	20	1	2	1	135	0	0	84.91%
pred. openc...	16	2	1	0	4	149	0	86.63%
pred. presta...	10	1	1	0	3	0	219	93.59%
class recall	95.97%	85.59%	82.32%	86.76%	78.49%	96.13%	96.05%	

Fig. 5. The confusion matrix of the Decision Tree model

3) *Naive Bayes model*

The Naive Bayes model was trained with an accuracy of 65.86%. Considering the model's confusion matrix (Fig. 6), it can be seen that the WordPress type is predicted with a high accuracy of 95.05%, but the prediction of the other classes is low, ranging from 0% to 10.59%. The results show that the WordPress type has a significant impact on the model predictions. As WordPress is the type of CMS relevant to the study, its removal would reduce the model's practicability.



	true wordpr...	true duda	true joomla	true other	true cmsma...	true opencart	true prestas...	class precisi...
pred. wordp...	2717	322	177	199	142	133	199	69.86%
pred. duda	19	2	1	3	2	1	4	6.25%
pred. joomla	13	0	0	1	2	1	4	0.00%
pred. other	26	5	0	6	1	2	0	15.00%
pred. cmsm...	52	5	8	6	18	2	1	19.57%
pred. openc...	16	1	1	1	3	4	5	12.90%
pred. presta...	46	5	8	3	2	10	15	16.85%
class recall	94.05%	0.59%	0.00%	2.74%	10.59%	2.61%	6.58%	

Fig. 6. The confusion matrix of the Naive Bayes model

Summarising the results, the K-Nearest Neighbours model had an accuracy of 69%, but the model was found to predict the WordPress type with 99.83% accuracy. Meanwhile, the other CMSs were not identified or predicted with low accuracy. This means that the predicted attribute types are not evenly distributed and the frequency of occurrence of WordPress is significantly higher compared to the other CMS attribute types.

The Naive Bayes model has an accuracy of 65%, but similarly to its K-Nearest Neighbours, classes other than WordPress had low prediction accuracy. Removing the WordPress class from the dataset would help to increase the prediction accuracy of the classes, but WordPress has a significant impact on the study as it is a universal and exceptionally popular CMS in Lithuania and worldwide. Removing the WordPress class from the dataset would make the trained model ineffective in predicting the CMS used by companies.

Using the Decision Tree algorithm with information gain and gain ratio criteria, the highest accuracies of 93.4% and 93.7% were achieved, with optimal depths of 21 and 37 respectively. Considering the results obtained, it can be concluded that the Decision Tree algorithm is effective in solving the task of classifying corporate CMS and predicting the type of CMS from corporate profile data.

Table 6
Highest model prediction accuracies and parameters

Classifier	Accuracy	Parameters
k-NN	68.5 %	k = 10
Decision Tree	93.5 %	information gain, depth = 21
Naïve Bayes	65.9 %	-

The results of the data analysis showed (Table 6) that the choice of an appropriate algorithm and the setting of its parameters are critical factors in determining the efficiency and accuracy of the final results.

5. Conclusion

The analysis of the studies shows that the articles use different research methods – regression, statistical analysis, etc. There is also a lack of research on the use of CMS by companies. Moreover, the proposed recommendations for CMS were based on analysis of CMS, usage statistics and surveys, but not on company profile data.

The selection of methodologies and tools for extracting company profile data and identifying CMS types has shown that the web browser and additional third-party CMS identification tools allow for the extraction of openly available information from online portals.

The analysis of the data mining models shows that applying the Decision Tree algorithm to the classification problem can achieve model accuracies greater than 90% in predicting companies' use of CMS based on company profile data.

The study used automatic construction of classification methods and developed five machine learning models - Naive Bayes, Decision Tree, Deep Learning, Logistic Regression and Random Forest - with similar accuracies ranging from 68.3% to 69.0%. The Naive Bayes model has the highest accuracy of 69.0%, while the Random Forest model has the lowest accuracy.

The results showed that the dataset is complex and the defined task is not easily solvable using automated model training. The performance of different algorithms depends on the dataset and the problem to be solved, so in order to develop an efficient model, k-NN, Decision Tree and Naive Bayes models were developed with individually defined parameters for each model. k-NN model achieved an accuracy of 69%, Naive Bayes model an accuracy of 65%.

Using the Decision Tree model with information gain and gain ratio criteria, the highest accuracies of 93.4% and 93.7% were achieved, with optimal depths of 21 and 37 respectively. In view of the results obtained, it can be concluded that the Decision Tree algorithm is effective in solving the task of classifying corporate CMS and predicting the type of CMS from corporate profile data.

References

[1] Fernandes S, Vidyasagar A. Digital Marketing and Wordpress. Digital Marketing and Wordpress Article in Indian Journal of Science and Technology 2015;8:61–8.

[2] Zujovic L, Kecojevic V, Bogunovic D. Application of a content management system for developing equipment safety training courses in surface mining. The Journal of the Southern African Institute of Mining and Metallurgy 2020;120.

[3] BuiltWith. CMS technologies Web Usage Distribution on the Entire Internet 2021. https://trends.builtwith.com/cms/traffic/Entire-Internet (accessed December 15, 2021).

[4] W3Tech. Usage Statistics and Market Share of Content Management Systems, April 2023 2023. https://w3techs.com/technologies/overview/content_management (accessed April 20, 2023).

[5] Skatikienė MM, Pauliukaitis D. Internetinių svetainių populiariausių turinio valdymo sistemų tyrimas. Pramonės inžinerija - 2017: jaunųjų mokslininkų konferencija, 2017 m gegužės 11 d: pranešimų medžiaga 2017:152–8.

[6] Hwang Y, Al-Arabiat M, Shin DH, Lee Y. Understanding information proactiveness and the content management system adoption in pre-implementation stage. Comput Human Behav 2016;64:515–23.

[7] Laumer S, Maier C, Weitzel T. Information quality, user satisfaction, and the manifestation of workarounds: a qualitative and quantitative study of enterprise content management system users. 2017;26:333–60.

[8] Drivas I, Kouis D, Kyriaki-Manessi D, Giannakopoulos G. Content Management Systems Performance and Compliance Assessment Based on a Data-Driven Search Engine Optimization Methodology. Information 2021, vol. 12, Page 259, 2021;12:259.

[9] Sodra. Atviri įmonių duomenys - atvira.sodra.lt 2023. https://atvira.sodra.lt/imones/rinkiniai/index.html (accessed May 6, 2023).

[10] Okredo. Lietuvos įmonės | Okredo 2023. https://okredo.com/lt-lt/ (accessed May 6, 2023).

[11] UAB "Verslo žinios." Rekvizitai.lt. Įmonių katalogas, įmonės 2023. https://rekvizitai.vz.lt/ (accessed May 6, 2023).

[12] UAB "Saulės Spektras." Info.lt - naujos kartos įmonių paieška! 2023. https://www.info.lt/lt (accessed May 6, 2023).

[13] Kaur P. Sentiment analysis using web scraping for live news data with machine learning algorithms. Mater Today Proc 2022;65:3333–41.

[14] Detect which CMS a site is using - What CMS? n.d. https://whatcms.org/ (accessed April 17, 2024).

[15] RapidMiner. RapidMiner Studio - RapidMiner Documentation 2024. https://docs.rapidminer.com/latest/studio/index.html (accessed April 22, 2024).

[16] Han J, Kamber M, Pei J. Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems) 2011.

[17] Arlot S, Celisse A. A survey of cross-validation procedures for model selection. 2010;4:40–79.

[18] RapidMiner. Auto Model - RapidMiner Documentation 2023. https://docs.rapidminer.com/9.4/studio/guided/auto-model/ (accessed May 13, 2023).

[19] Cássia Sampaio, David Landup, Dimitrije Stamenic. Guide to the K-Nearest Neighbors Algorithm in Python and Scikit-Learn 2023. https://stackabuse.com/k-nearest-neighbors-algorithm-in-python-and-scikit-learn/ (accessed May 1, 2023).

[20] Hastie T, Tibshirani R, Friedman J. Springer Series in Statistics the Elements of Statistical Learning Data Mining, Inference, and Prediction 2017.

[21] Daniyal Shahrokhian. The Naive Bayes Algorithm in Python with Scikit-Learn 2018. https://stackabuse.com/the-naive-bayes-algorithm-in-python-with-scikit-learn/ (accessed May 3, 2023).