A CRISP-DM and Predictive Analytics Framework for Enhanced Decision-Making in Research Information Management Systems

Otmane Azeroual [1,4*], Radka Nacheva [2], Anastasija Nikiforova [3], Uta Störl [4]
[1] German Centre for Higher Education Research and Science Studies (DZHW), Schützenstraße 6A, 10117 Berlin, Germany
[2] Department of Informatics, University of Economics - Varna, 9002 Varna, Bulgaria, 77 Knyaz Boris I blvd.
[3] Institute of Computer Science, University of Tartu, 51009 Tartu, Estonia, Narva mnt 18
[4] University of Hagen, Universitätsstraße 47, 58097 Hagen, Germany
E-mail: azeroual@dzhw.eu, r.nacheva@ue-varna.bg, nikiforova.anastasija@gmail.com, uta.stoerl@fernuni-hagen.de
* Corresponding author

Keywords: predictive analytics (PA), current research information systems (CRIS), research management, machine learning, topic modeling, decision-making

Received: December 29, 2023

The age of digitization has led to a significant increase in the amount and variety of data, particularly within the research domain, where data previously stored in paper form has now been digitized and integrated into research management processes. The rapid growth of Big Data, driven by technologies like the Internet of Things, presents challenges for conventional data processing methods. However, data alone, stored in silos, lacks value. To unlock its potential, data must be analysed and processed to generate insights and predictions that enable evidence-based decision-making. Predictive Analytics (PA) is a powerful tool for this purpose. By leveraging PA and advanced statistical methods, predictive models for research management can be developed, helping to forecast research trends and outcomes, which in turn, provides decision-makers with a reliable, forward-looking basis for strategic decisions in research management. This paper explores the application of PA in Current Research Information Systems (CRIS) to enhance decision-making. A case study using metadata from 20,000 publications indexed in Scopus demonstrates how PA can identify emerging research topics and predict future trends. Machine learning algorithms such as Support Vector Machine (SVM), k-Nearest Neighbor (kNN), Random Forest, and Tree classifiers were employed, with metrics such as Area Under ROC (AUC), classification accuracy (CA), F1-score, precision, and recall evaluated. The results indicate that the kNN algorithm provided the highest performance with an AUC of 0.451 and a classification accuracy of 87.4%. These results show that predictive models can reveal significant patterns in research data, supporting data-driven decision-making for research management. Additionally, the study applied Latent Semantic Indexing (LSI) and clustering techniques to identify and categorize key topics within the data, showing a thematic focus on areas such as smart cities and urban intelligence before predictions, and CRIS applications after predictions. The findings illustrate how PA can optimize research management by identifying gaps in research and forecasting emerging topics, thereby aiding institutions in making more informed, evidence-based decisions.

Povzetek: Predstavljen je okvir CRISP-DM s prediktivno analitiko za CRIS sisteme, ki s strojno učno klasifikacijo in tematskim modeliranjem omogoča boljše napovedovanje raziskovalnih trendov in podporo odločanju.

1 Introduction

Research management in research institutions is in a constant process of change seeking compliance with the state of the art and, preferably resilience and sustainability. Initially, research data typically referred to as research information were processed in IT systems to ensure efficient data processing. Research information and data in general are usually compared to oil [1], which can, however, be seen as an overly simplistic and highly inaccurate comparison, as Forbes points out [2], according to which the only similarity between data and oil lies in their "power", including the power of those, who own them - Forbes compares data owners like Alibaba, Google, Twitter, Facebook etc. to oil barons. Otherwise, a deeper comparative analysis shows that oil and data are very different, e.g.: (1) oil is a finite resource, while data are not - instead, data are effectively infinitely durable and reusable, while treating them like oil, i.e. storing in siloes, reduce their value, usefulness and potential in general; (2) oil requires huge amounts of resources to be transported to where and when it is needed, while for data – they can be replicated indefinitely and moved around at very high speeds and low costs; (3) when oil is used, its energy is being lost (as heat or light), or permanently converted into another form such as plastic, while the usefulness and utility of data tend to increase with their use, i.e. new uses arose, including but not limited to the use of the data as

training data; (4) as the world's oil reserves dwindle, its extraction becomes increasingly complex and expensive, while for data – they become more available, including but not limited to the technological advances, as well as due increasing number of data producers; (5) oil drilling is associated with environmental damage and exploitation of finite natural resources, while data mining does not – at least there is no intrinsic environmental damage and exploitation of limited and finite natural resources (except for the electricity used to run system and the relatively low trend of "green computing" with reference to "sustainable computing" for their further processing). To put it simply, Forbes is suggesting that we stop comparing data to oil and stress that if data are expected to be discussed as a source of energy or fuel, it makes much more sense to compare them to renewable sources such as solar, wind and tides. All in all, data can be seen more than oil [2, 3]. A trend that is primarily based on the collection and processing of large amounts/volumes of internal and external data and, ideally, on the creation of added value and gaining competitive advantages. With globalization and increasing technological progress, the volume (and at times also variety) of data that are produced/ generated, collected, transferred, processed and analysed is rapidly increasing. Institutions often underutilize the potential that comes from the increasing amounts of data [4, 5].

Ensuring the usability of the growing volumes of data is critical to competition and provides crucial and decisive growth drivers not only in business but also in research. To do this, research information must be processed, analysed and used. Data analysis can be very time-consuming and often only reflects the past due to poor data management, ignorance of available data that have been collected in the past (e.g., as part of previous research projects) and time spent collecting or even re-collecting data, compiling them, etc. In addition, increasingly complex and dynamic factors in the environment of all organizations lead to growing uncertainty about future developments. This poses new challenges for research management and requires new data-driven planning and control instruments that not only look into the past but above all into the future [6].

Research decision-makers who specialize in analysing research information and predicting/ forecasting future events more accurately gain a better understanding of their current and future scientific environment through the knowledge gained [7]. This, in turn, enables more meaningful and informed decisions.

The process of analysing large heterogeneous amounts/ volumes of data to forecast and plan future developments is associated with so-called "predictive analytics" (PA) seen as part of Business Intelligence (BI) [8]. Recent studies show that PA is increasingly being used in companies - predominantly large companies, where, according to [9], in 2020, of the 52% of companies worldwide that use PA, 65% of companies using PA have between $100 million and $500 million in revenue, with significantly lower adoption in SME.

However, the use of PA in current research information systems (CRIS) (also known as Research Information Systems (RIS) and Research Information Management Systems (RIMS)) is still in its infancy, and some numerous challenges and risks prevent research organizations from the PA implementation. Therefore, the potential use of PA is currently being discussed by institutions, practitioners, research advisors and scientists in professional circles with limited literature and developments in this area [10, 11, 12, 13], mostly remaining of a theoretical nature. At the same time, it is appropriate to consider scenarios for applying PA in CRIS, to explore the resulting potential benefits in various research management processes.

In this regard, **our paper aims to propose** a predictive analytics-based decision-making model for IT-supported research management. First, the PA instrument will be outlined and examined, followed by an overview of the current state of PA research in IT-supported research management to unlock the (unused) potential of PA. Finally, the implications, challenges, and risks of implementing PA in IT-supported research management will be identified and evaluated.

Therefore, the main research question (RQ) that we define and seek to answer in this study is: What potential benefits arise from PA in IT-supported research management (such as CRIS) and decision support in institutions in general? We then provide a step-by-step guide to the use of the proposed model, demonstrating its applicability in the chosen domain/area of application - CRIS. Finally, we define possible future directions.

2 Background

Numerous strategic decisions are made every day, which are expected to be based on the data held by the actors of the respective organization or entity, be it a company, its department, an individual, a university, or a research center. Decision-making is part of every entity from the very beginning of data collection, whether internal, external, or both [14]. In recent years, science and practice have postulated that with new data sources and changing usage behavior, new opportunities are opening up for institutions to analyze research information to ultimately be able to make faster and better strategic decisions. In this regard, data is one of the most important resources for the future of institutions, and PA is part of business analytics and business intelligence, with many touchpoints to data

Table 1: Summary of Related Approaches to PA

Study	Methods Used	Application Domain	Quantitative Results	Identified Limitations
Piryonesi & El-Diraby [17]	Predictive Analytics (PA), machine learning, data mining	Construction industry	Improved predictions of project delays and costs	Limited to construction, lacks generalization to other domains
Spencer [18] & Akter & Wamba [19]	Machine learning, data mining, recommendation systems	E-commerce, Retail	Increased sales, improved customer targeting	Data privacy concerns, overfitting in models
Tanlamai et al. [20]	PA, predictive modeling	Retail and inventory management	Reduced overstocks and understocks, saved costs	Does not account for market volatility
Azeroual et al. [13]	Sentiment analysis, PA	Social Media and Technology	Sentiment-based decisions improved customer engagement	Dependent on quality of social media data
Balbin et al. [21]	PA, open data, smart city applications	Smart cities, transportation services	Optimized traffic flow, reduced congestion	Data reliability, integration challenges

mining and machine learning [15]. These and other statistical methods make it possible to predict the likelihood of future events. While data mining involves examining large amounts of data to identify patterns, trends, and relationships, machine learning involves algorithms that independently acquire new knowledge and generate data models used for forecasts and decision-making [16].

To examine the central research question regarding data processing in CRIS and the PA process, it is helpful to look at the work of previous researchers. The table below summarizes key methods, domains of application, and results of studies that have explored predictive analytics in various fields. This comparison allows us to highlight the novelty and relevance of our approach for more efficient research management.

As shown in Table 1, existing research has primarily applied PA in specific industries such as retail, e-commerce, construction, and social media. However, these approaches have limitations, such as narrow application domains or a focus on specific data types (e.g., customer data in e-commerce or traffic data in smart cities). For example, while in [17] authors are focused on improving project cost and schedule predictions in the construction industry, their methods are not directly applicable to the broader context of research management. Similarly, while PA in retail has shown promise for inventory management, the methodologies often fail to account for more complex, multidimensional research environments.

The novelty of the presented work lies in applying predictive analytics specifically to research management using CRIS systems. Unlike traditional approaches in business or industry, this study explores how predictive models can optimize research performance and decision-making by utilizing large-scale research data. Furthermore, by integrating data mining techniques and machine learning into CRIS, we address existing gaps such as the need for more robust, scalable models that can handle the dynamic and heterogeneous nature of research data.

2.1 Research management

Research planning and control, and the associated research reporting obligations, have a long tradition [22]. Although initially, the government/ state ministerial bureaucracy in the context of the development of the first approaches to the control of research / scientific policy was the main demand for research information, the transition to a model of governance of the self-regulatory higher education sector in the 1980s was accompanied by a change in the need for information on research performance, both qualitative and quantitative [23]. At the heart of these changed information needs is the new mission of the university as an organization. In the course of this reformist discourse, the role of company management and the administration of research organizations is strengthened and differentiation of professional roles and an increase of professionalism of the administration and management of the university can be observed [24]. This not only increases the need for information for the goal-oriented direction of university management and control. In addition, given the

increasingly multidimensional understanding of the quality of research, the requirements and demands on the complexity and information density of research information are increasing.

Research information or information about research serves the interests of a wide range of data users, including research organizations themselves, but also research funders, foundations, companies and other stakeholders. This is also due to digitization, namely because the way researchers work has changed and more and more data is being digitized and stored electronically instead of in paper archives [5]. Accordingly, practices of electronic/digitized data processing and further use are gaining importance, among which research information systems such as CRIS play a prominent role. As information systems, they support the exchange of data and the networking of research information from different sources and are suitable for strategic research reporting with a variety of output and analysis functions [25]. According to [26], CRIS represents research management information and provides the ability to merge corporate and academic research activities, reduce duplicate data entry, increase data quality, identify authoritative sources of information and understand complex relationships between researchers, projects and outcomes [27].

Currently, many academic research institutions in Europe are involved in software-based research management projects (e.g., introducing CRIS) that offers a suitable solution. CRIS supports the entire research process in academic institutions and promotes transparent, uniform and structured documentation of research. For example, CRIS should be able to link projects with funding and research results and enable evaluation and assessment within an institution as well as a comparison with other institutions [26]. In addition, CRIS can be used to manage research projects, research results, research resources and research funding. This, in turn, reduces response times and coordination paths.

2.2 Predictive analytics (PA)

PA is a relatively new phenomenon - companies from different industries use this method to optimize business processes, remain competitive and act flexibly. Measuring possible future events is important for banking and insurance, retail and e-commerce, among others [28].

The literature has shown that PA is a subset of data analysis [17, 29]. PA is a corporate planning tool and describes mathematical-statistical processes that recognize relationships in data and, based on this, predict the probabilities of future events, developments and trends. The analysis tools used, such as self-learning algorithms (machine learning), statistical models and methods are summarized under the term data mining. In addition, other methods for processing and generating data are used, such as Text mining, in which unstructured text information (e.g. articles or blogs) is analysed. The future forecasts prepared by PA are more accurate than conventional forecasts.

Data-based analysis, planning and control have always been a central part of research management. However, the increasing availability of internal and external amounts of large research information poses major challenges for CRIS. Modern data analysis techniques are therefore becoming more and more relevant and offer more and more possible applications. However, this increase in data volume does not automatically mean that added value is generated for the institutions. Only through the use of modern data analysis techniques can additional insights and information be gained from large amounts of data. Conventional analysis techniques often cannot keep up with the large amounts of data, which is why modern approaches such as PA are required to deal with the amounts of data.

According to [30], forecasts in the context of PA play a particularly important role in the risk management of financial service providers. For example, banks can use the credit rating to calculate the risk that future instalment payments cannot be made. Among other things, insurance companies can forecast future damage and possible costs and thus determine suitable tariffs. However, the detection of fraud is also an area of application.

According to [20], replenishment in retail depends on PA for needs-based purchase of goods. With precise forecasts, overstocks and understocks in the warehouse can be avoided. In this way, costs can be saved that would arise from short sales and the availability of the range is always guaranteed. Accurate sales forecasts are particularly important for fresh and perishable products. Influencing factors are special offers, public holidays and the weather.

According to [18] and [19], e-commerce can derive future purchase interests from collected customer data, such as information on products already purchased, and thus the probability of successful cross-selling or up-selling. A well-known example is purchase recommendations in an online shop. Based on the products purchased, similar products are identified that the customer may also be interested in.

In addition to these above areas, PA can also be used in other areas and generally tries to answer the question: "*What will happen?*", because digitization is progressing rapidly worldwide. These other examples include also the application of PA not only in business, includes predicting stock prices [31], management accounting [32] or medical/ healthcare domain, including, disease control and prevention [33], and predicting hospital costs and their savings [34], but also in education, i.e., to predict performance in MOOCs [35]. Yet another area, in which PA proved to be highly beneficial is crisis management regarding both health-related crises and natural disasters management, e.g., as a tool for effective fire risk management by analyzing fire incidents' data from the National Fire Incident Reporting System (NFIRS) and weather data [36]. Balbin et al however, did so concerning the topic of Smart Cities applying PA to support smart transportation services using open data [21]. In [37] also mentions industrial use-cases such as energy to forecast supply and demand, and predict the impact of equipment costs, downtimes / outages, and other variables, aerospace

to predict the impact of specific maintenance operations on aircraft reliability, fuel use, and uptime, while the biggest airlines - to predict travel patterns, setting ticket prices and flight schedules as well as predict the impact of, e.g., price changes, policy changes, and cancellations, and, of course, retail as a subset of business we mentioned above, where PA allows retailers to follow customers in real-time, delivering targeted marketing and incentives, forecast inventory requirements, and configure their website (or store) to increase sales. PA are also widely used in the business process management area giving rise to what is called predictive process monitoring (PPM) [38, 39, 40]. In recent years, PA of social media data has attracted considerable attention in both the research community and the business world due to the essential and actionable information it can provide [41]. In 2023, an intelligent PA decision framework was developed by authors [13], which was tested through sentiment analysis on Twitter data. It was shown whether and how PA can help support data-driven decisions, including technological developments. Digital processes in institutions generate huge amounts of data that are growing every day. Industry 4.0 and the ongoing transition to Industry 5.0 and the Internet of Things (IoT) are further accelerating this development. To run a company successfully and make the right decisions, detailed and deep insights into the many different processes are necessary. Companies face the challenge of aggregating, managing and evaluating a large amount of data.

To facilitate the rapid and efficient development of a PA model, a systematic approach is essential [42]. In a survey conducted by [43], a panel of experts emphasized the need for advanced PA skills, along with an artistic and creative mindset, for successful model building. However, only 15% of the respondents reported using the CRISP-DM methodology for implementing PA [43]. CRISP-DM is a widely used, cross-industry approach to data mining that was conceived in 1996 as part of an EU-funded project and is an acronym for Cross-Industry Standard Process for Data Mining [44]. Given its wide popularity and proven benefits when applied in real-world scenarios, we will use it as a reference model for our predictive analytics-based decision-making model for IT-enabled research management.

Therefore, let us briefly discuss CRISP-DM process model (Figure 1), which defines six basic phases we will follow when developing our PA model [44, 45]:

- **definition of the project scope:** determination of the desired goals and results as well as the derivation of the rough procedure;
- **data exploration**: analysis of the raw data to determine the most appropriate data and models, as well as identification of any problems (e.g., anomalies and other data quality issues);
- **data preparation**: construction of the final data set by selecting, extracting and transforming the data;
- **construction of the model**: creation of models with suitable statistical methods and consideration of the defined requirements. Different statistical methods can be used to create a PA model. Results from a survey by [46] showed that linear regression and decision trees are the most commonly used. Cluster analysis, time series models and logistic regression follow in the next places;
- **evaluation of the model**: verification, validation and optimization of different models by comparing the requirements;
- **use and management of the model:** integration and application in business processes and decisions. Further development and optimization of the model, e.g., accuracy improvement or reduction of redundant processes.

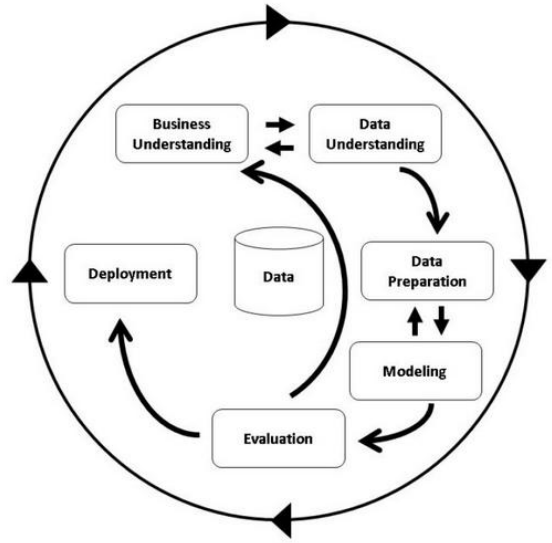

Figure 1: CRISP-DM Model [45].

Analysis of research information is essential for the successful realization of a CRIS project. Big research information refers to the scientifically meaningful acquisition and use of decision-relevant knowledge from qualitatively diverse and differently structured information that is subject to rapid change and is generated in an unprecedented volume. Compared to conventional evaluations, statistical methods (linear regression, decision trees, cluster analysis, time series models and logistic regression) also play a decisive role in PA. Automated analyzes and intelligent algorithms can process large amounts of data. Internal research information can be supplemented with external and previously unavailable data to analyze relationships and refine predictions.

3 Opportunities and challenges of PA in CRIS

As with any other system and the incorporation of technological support, there can be advantages and disadvantages, as well as opportunities and challenges associated with the PA use [47]. In addition, due to the importance of large amounts of research information and increasing volumes of internal and external data, new job profiles related to data management are emerging. Let us discuss them in more detail.

One of the main advantages of PA is the faster and more efficient generation of forecasts for business planning and control [48]. Thanks to automated solutions, this process can be carried out more often and, above all, with lower personnel costs. The inclusion of new internal and external data sources also positively impacts the quality and reliability of these forecasts, which can lead to savings and cost optimization opportunities across departments [49]. Forecasts can be used to take proactive action at an early stage to counteract undesirable consequences or changes in a targeted manner. Other benefits mentioned concerning PA are interdepartmental transparency in research institutions, increased accessibility and satisfaction, and tailored products and services for researchers and decision-makers [46, 50].

About the research results, institutions show that a CRIS employee spends around two-thirds of the time in the research process on manual activities that do not generate direct added value [51]. The reasons are varied and range from a lack of automation, and unstructured research information, but also increasing complexity of data processing. It has been estimated that the degree of impact of digitization correlates with the resource expenditure, i.e. where it finds its most influential position, most of the CRIS employee capacities are accumulated. These are central research processes and reporting, whereas envisaged technologies for this are Big Data, Robotic Process Automation, Machine Learning and Predictive Analytics. But these are also technologies that will permanently change the role model of the research manager or CRIS employee.

In addition to the advantages and opportunities of CRIS, the handling and integration of additional data sources and data volumes entails various risks. For example, interfaces to all internal and external data sources must be maintained and research information must be constantly monitored and validated to enable data identification and integration. Caution is required, particularly with regard to the monitoring and validation of research information, including but not limited to data quality management, which includes not only completeness checking but also checking the plausibility, timeliness, accuracy, consistency and availability of the research information. However, such data sources also lead to enormous data growth and increasing demands on research management that integrates research information from different sources into CRIS. In addition, the collection, storage, deletion, modification and disclosure of certain research information involve other risks related to research information compliance with privacy policies and legal requirements. Depending on the area of activity, this includes national and international regulations that must be checked in advance and always observed. On the other hand, moral and ethical aspects must not be ignored. By creating transparency, negative effects on reputation can be prevented. All this is expected to be done through ***data governance***, which is expected to become an integral part of the data management environment, linking business policy with data management and forming a regulatory framework for the above [52].

As a result, in the course of new activities, technologies and the associated challenges, new job profiles have recently emerged in order to overcome the new hurdles [53]. These include the *data architect*, *database specialist*, *data scientist*, and *data steward*, each of whom deals with additional volumes of data in one way or another. The *data architect* and the *database specialist* take care of the management of the research information. The *data architect* is responsible for the institution-wide data architecture and, based on the business models, derives where and how research information is provided. The *database specialist* deals with topics such as database system technology (CRIS), distribution, archiving, reorganization and recovery of research information. A *data scientist* specializes in the field of business analytics and deals with data analysis and data interpretation, which methods and tools of data mining, statistics and visualization of multidimensional relationships among data masters and use to extract unknown facts and patterns from data and future forecasts derive [54].

4 CRIS and business analytics & Intelligence models

The data universities and research institutions deal with can be seen as a “data lake” (or a “data swamp” if a proper data governance mechanism is not in place), which is fed from a wide variety of data sources that are constantly increasing. These data sources include universities, research organizations, researchers, teaching staff, project managers, funding organizations, publishers, libraries, decision-makers, media, the general public, intermediaries/brokers, and enterprises [55] (Figure 2). This vast amount of structured, unstructured and semi-structured big research data or information is referred to as big data. But big data without PA is like a car without an engine. From clarifying the past to predicting researcher behaviour and deriving promising decisions and activities in the research process, big data technologies become a sustainable competitive advantage.

A more specific variation of this type of system (CRIS) is a system using the Common European Research Information Format (CERIF). CERIF is a conceptual model for the domain of scientific research, where systems that use it can also be used to access and review financial, human resource and project management information of an organization [56]. One of the great advantages of CERIF from the point of view of organizations is that the use of these systems is related to the processing of administrative data in business intelligence strategic decision-making activities [56].

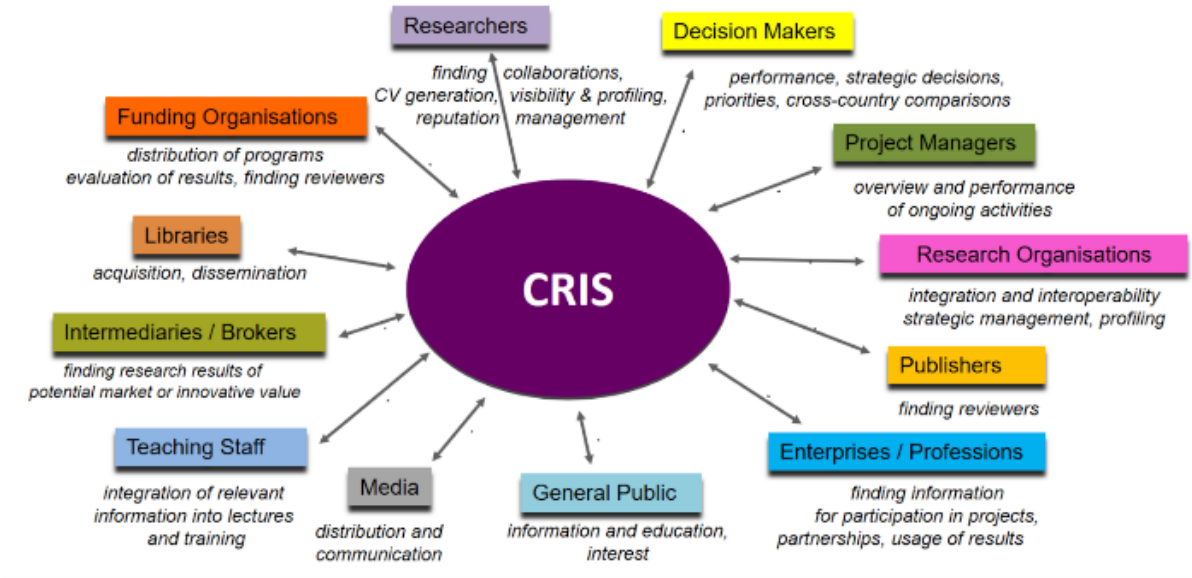

Figure 2: Research information origin [55].

CERIF are useful in creating comprehensive analytics to make informed strategic decisions on resource allocation and target setting [57]. Research Information Infrastructure on the institutional level is shown in Figure 3. This type of system receives data from multiple institutional and external sources, including institutional administrative systems, research databases, Internet. Information can be structured (spreadsheets and documents), semi-structured, or completely unstructured (texts and multimedia files from websites extracted using data mining).

As stated in [57], they are useful when dealing with multiple data sources that are not related to each other and that have different identifiers. This leads to complicated research management, slow decision-making and missed opportunities.

PA enables predictions about future events and trends based on complex data analysis. In addition to classic data mining methods such as cluster analysis, regression analysis and association analysis, additional methods and processes such as machine learning or text mining are used to evaluate unstructured text information from documents, e-mails or blogs. According to Gartner's Analytics Maturity Model, PA is the third evolutionary stage of analytics methods (see Figure 4). The first two phases, namely, descriptive analytics and diagnostics analytics, describe and explain the past based on the analysis of historical data only. Phases 3 and 4 - PA and prescriptive analytics - enable predictions and automate decisions to affect them.

The use of analytics applications in the context of CRIS can offer real added value for CRIS employees. While existing analytics applications are limited to the description (descriptive analytics) and explanation of

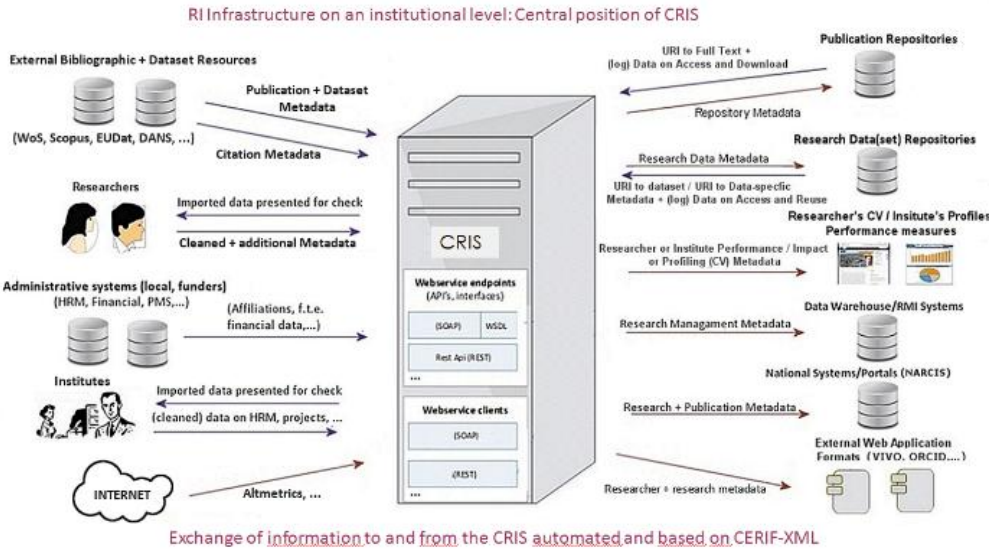

Figure 3: Research information infrastructure on institutional level: central position of CRIS [56].

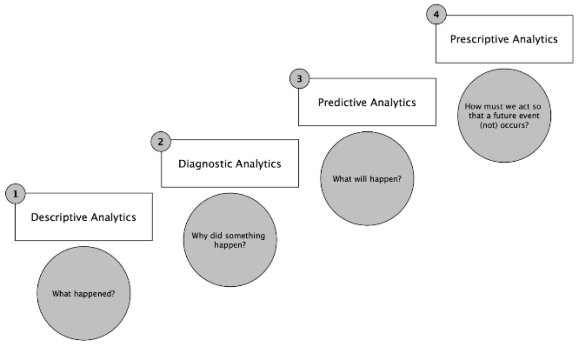

Figure 4: Gartner analytics maturity model [58].

business processes (diagnostic analytics), the development is moving in the direction of forward-looking (PA) and recommending preventive analyses (prescriptive analytics). While PA creates forecasts about what is happening or could happen, prescriptive analytics provides recommendations for action based on these forecasts to ensure this future event occurs or doesn't occur. Prescriptive Analytics allows the CRIS user to understand how different actions affect the outcome of the research process and thus provides information for the most optimal decision.

PA requires significant, if not to say, huge amounts of data. However, PA comes with a number of prerequisites, including but not limited to research information and data quality (see Figure 5).

Data quality is crucial for advanced analysis methods. Poor data quality in Analytics leads to poor performance. There are other requirements for PA and research information underlying these analytics. First, research information and data are certainly the limiting basis of any PA method, i.e., the better the data quality, the better the forecast. However, the research information must also be relevant and should therefore first define the fundamental problem to then decide which research information is crucial and can add value to the analysis. A meaningful CRIS is therefore essential to be able to access relevant research information quickly and easily. Another focus of the PA workflow is the processing of research information. Secondly, the facility culture and the internal infrastructure are often decisive for the software used, i.e., how (is the amount of) data managed? Is there a big database like CRIS or self-developed / ad-hoc databases or information systems? Are PA algorithms developed independently or is ready-made software used, into which the pre-processed research information only has to be integrated? Finally, CRIS employees not only need to understand the business problem but also need a team of

Figure 5: Steps for providing data with assured quality.

specialists who bring their expertise to bear in different areas. This, in turn, requires people who are knowledgeable about research information and who can handle the management and storage of research information. Equally necessary are employees who can understand, apply and interpret the algorithms behind PA. The range thus extends from data engineer to data analyst and data scientist. Depending on the size of the facility, several teams may even be necessary. But of course, the presence of knowledgeable people is not the only component, which is important for the above, where proper data governance is not of less importance as a prerequisite for the successful fulfilment of the set tasks by the above actors.

The benefits of PA are diverse and give facilities more than just a glimpse into the future. A PA process or approach consists of a series of steps that build on each other. The entire process is iterative, i.e., the desired result is approached in several passes; the steps can therefore be repeated:

- **Step 1 – setting objectives** - objectives of the analysis for the institutions are defined. The data sources and their availability are checked;
- **Step 2 – data collection** - the unified data collection from different data sources takes place;
- **Step 3 – data review and processing:** data quality and reliability of data delivery are crucial for any data-driven activity, i.e., research information. After the data has been collected/obtained, the data is checked and cleaned and the research information is made available for upcoming analyses.
- **Step 4 – building predictive models**, where the future forecasts can either relate to points in time and values (such as sales forecasts for a defined point in time) or to classifications (such as risk analysis). Analyzing for a risk means dealing with a possible incident in advance. This incident should be realistically assessed for the future.
- **Step 5 – model test and adjustment**, when the models are fed with the research information and tested, and corrections, if there is a need for such, are made;
- **Step 6 – provision for integration into facility processes**, when findings can be used for making decisions in the facilities.

5 Method

In this section, we apply a data mining research methodology to analyze the metadata of scientific papers. focusing on the application of CRIS in the context of scientific research classification. The approach is structured into well-defined phases, ensuring a robust and replicable research framework.

5.1 Overview of methodology

The study utilizes a random sample of 20,000 English-language paper metadata retrieved from the Scopus indexing database, spanning multiple journals between July 1st and July 11th, 2023. The search included key

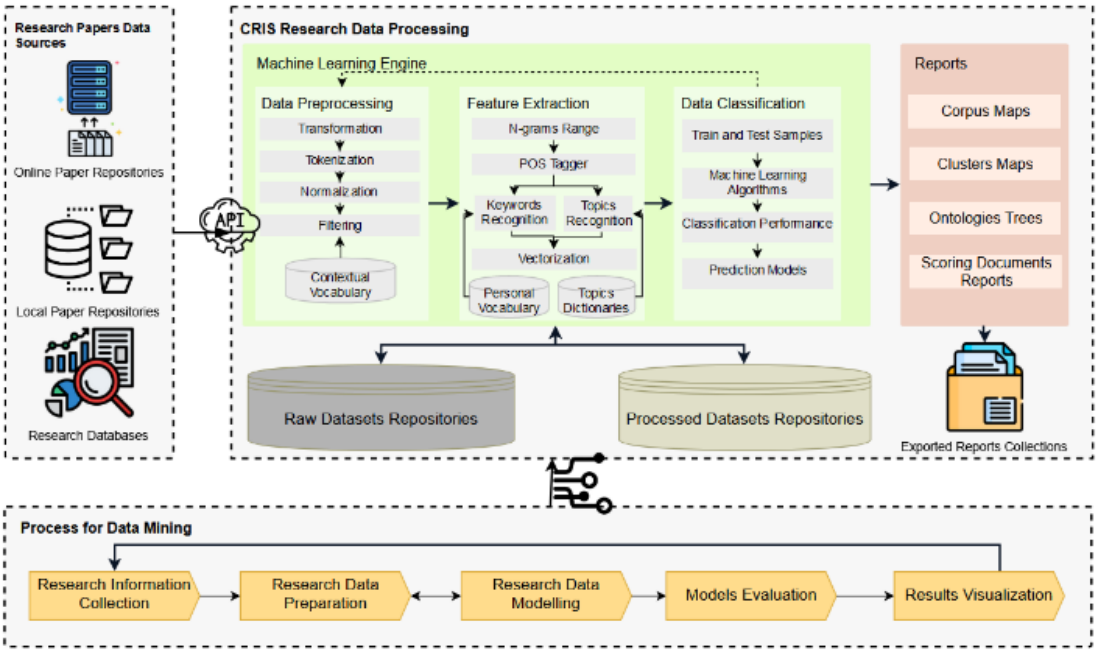

Figure 6: Proposed PA decision-making model.

phrases such as “smart cities”, “intelligent cities”, “cities IQ”, "CRIS development," "CRIS management," and "CRIS systems”, among others. The metadata includes titles, authors, affiliations, abstracts, keywords, publication years, and citation counts.

This methodology is tested with a specific use case, We selected the “smart cities” topic because CRIS platforms are crucial for their data management, knowledge dissemination, interdisciplinary collaboration, and real-time monitoring [59]. They collect, curate, and integrate diverse datasets from various sources, facilitating the integration of data for city planning and decision-making. Smart cities use technology, IoT, and data to optimize urban systems and improve quality of life. CRIS supports smart cities by providing a Knowledge Repository, enabling data integration, and decision support. Intelligent cities integrate advanced technologies with knowledge and innovation ecosystems to foster economic growth and creativity. CRIS contributes by interdisciplinary collaboration, dynamic knowledge sharing, and capacity building. CRIS aids in this by providing data benchmarking, research-driven metrics, and progress tracking. CRIS development is linked to cities' needs by customizing for urban applications, enhancing interoperability, promoting open science, and incorporating AI, machine learning, and predictive analytics. The growth and evolution of CRIS systems are inherently linked to their application in urban contexts, driving their development to meet cities' needs.

The independent variables in this analysis include the type of publications (Article, Book, Book Chapter, Editorial, Erratum, Review, Short Survey) and the number of citations, while the dependent variables are the recognized keywords, topics, and ontologies. The results of this experiment are presented in the following section 5.

For this study, we employ the CRIS-integrated PA decision-making model for scientific research classification (Figure 6). This model is based on the standard CRISP-DM phases, in conjunction with our previous research [26, 60, 61], as well as research conducted by [40, 62, 63, 64].

The data mining process on which our proposed model is based consists of 5 phases, each of which corresponds to specific activities involving external sources or the toolkit available to the CRIS. The first phase of the model is associated with obtaining data for processing (phase #1: collection of research information). The remaining phases are based on the computational capabilities of CRIS for processing scientific research data - full texts or metadata. The CRIS "Machine Learning Engine" subsystem is used in the next three stages of the data mining process. The CRIS “Reports” subsystem is used at the last stage of the process of outputting the research results (phase #5: results visualization). In other words, the process consists of:

The data mining process on which our proposed model is based consists of 5 stages, each of which corresponds to specific activities involving external sources or tools available to the CDIP. The first stage of the model is associated with obtaining data for processing (collection of research information). The remaining stages are based on the computing capabilities of CIRI for processing scientific research data - full texts or metadata. The CRIS "Machine Learning Engine" subsystem is used in the next three stages of the data mining process. The CIRI “Reports” subsystem is used at the last stage of the process of outputting research results. In other words, the process consists of:

Phase #1: collection of research information, which relates to retrieving scientific papers from several sources. There are three main types of data sources for processing by CRIS – Internet repositories, local repositories and

research databases. Datasets can be retrieved in the form of metadata in structured SQL, XLS, YAML, XML format or full-text papers in pdf, doc or html format. Online repositories and research databases usually have APIs available to facilitate the retrieval of scientific papers. On the other hand, databases such as Scopus and WoS have tools for exporting metadata in CSV, RIS, BibTeX, TXT format. CSV format contains a set of labels and columns, i.e. metadata is exported in a structured form and can be used when performing classifications of scientific papers. They can be downloaded and processed in real-time, or downloaded to a local repository and then processed by CRIS tools;

Phase #2: research data preparation, which relates to the processing activities of the extracted datasets. By default, the data mining process transforms the input data, such as converting all letters to lowercase, removing accents, detecting HTML tags, and parsing out text only. URLs are removed from the text. A tokenization method is chosen for the transformed text, which can separate words according to whitespace, punctuation, or a regular expression. The preprocessing further includes stemming and lemmatization applied to words depending on the language of the papers. For instance, in English texts, the WordNet Lemmatizer is used to convert words like "running" into "run". Finally, the dataset is filtered by removing or keeping words based on previously prepared dictionaries. The dataset characteristics are also examined, including the class distribution, to ensure the dataset is balanced for effective training. The preprocessing steps ensure that the data is ready for feature extraction in modeling the research data;

Phase #3: research data modelling, where feature extraction of the processed dataset is performed most often by creating n-grams from tokens or running part-of-speech tagging on tokens. Depending on the objectives of the study and the type of data processed, keyword and topic recognition techniques and algorithms, as well as vectorization, can also be applied. Personal vocabularies or predefined topic dictionaries can be used here as well;

Phase #4: models' evaluation – The ultimate task of the CRIS Machine Learning Engine subsystem is to evaluate a model by applying machine learning algorithms to be used to create prediction models. In the first place, the training and test datasets are formed. Then the machine learning algorithms are chosen depending on the type of processed data and the objectives of the study. We selected algorithms such as Support Vector Machine (SVM), k-Nearest Neighbor (kNN), Decision Tree, and Random Forest due to their ability to handle complex, high-dimensional datasets. SVM is particularly effective for margin-based classification, while Random Forest provides robustness against overfitting. kNN is ideal for simple tasks where non-linearity is not prominent. The Decision Tree model is intuitive and works well for interpretability. After that, an assessment of the performance and adequacy of the applied algorithms is carried out according to metrics specific to the type of training. Another option is to apply the K-means with outlier removal (KMOR) algorithm is a variant of the standard K-means algorithm, incorporating an outlier cluster (k+1) to account for non-fitting objects [65]. The performance is evaluated using a variety of metrics including AUC, Classification Accuracy (CA), Precision, Recall, F1-Score, and Matthews Correlation Coefficient (MCC) to provide a well-rounded view of the model's performance. Finally, the prediction model is formed. Depending on the type of data, if necessary, predictions can go through text processing again to correctly form the final results of the experiment. On this basis, the subsequent activities are carried out in the final phase of the data mining process. These can be predictions of trends in topics, categories, ontologies, types of publications, impact of papers based on citations, and others according to the objectives of the research;

Phase #5: results visualization is the final phase of the model that is related to the visualization of the results through reports. They can be in the form of texts, tables, graphs or figures. They can visualize the distribution of datasets in clusters, categories, trees, as well as show statistics on the scoring of documents, ontologies, distribution of datasets by topics, etc. according to the type of data.

The proposed model is domain-agnostic and can be applied across various sectors due to its general nature. To demonstrate its functionality and conduct testing, the model is applied to a selected dataset, providing a step-by-step guide to its implementation.

Hyperparameter tuning is performed using Grid Search in combination with Cross-Validation. This process optimizes the parameters for each algorithm, such as the kernel type in SVM (linear, RBF), the number of trees in the Random Forest, and the depth of the decision tree. These tuning strategies ensure that the model achieves optimal performance. Hyperparameter tuning is crucial for optimizing algorithm performance. Common strategies include grid search, random search, Bayesian optimization, gradient-based optimization, evolutionary algorithms, hyperband, and cross-validation-based tuning [66]. Grid search is computationally expensive, random search is faster, Bayesian optimization balances exploration and exploitation, gradient-based optimization uses gradient descent, evolutionary algorithms refine hyperparameter combinations, and hyperband allocates resources efficiently.

5.2 Ensuring replicability

Replicability is a fundamental principle of scientific research, enabling other researchers to validate, extend, or adapt the methodology. To achieve replicability, the study provides a detailed account of the tools, libraries, and processes used throughout the research.

The implementation of the methodology is grounded in Python (v3.x), which supports a comprehensive ecosystem of libraries for data science. Key components include:

- **Data Processing**: Pandas (v2.2.3), NumPy (v2.2.1).
- **Text Preprocessing**: NLTK (v3.12), SpaCy (v3.7), and regular expressions (re library - regex 2024.11.6).

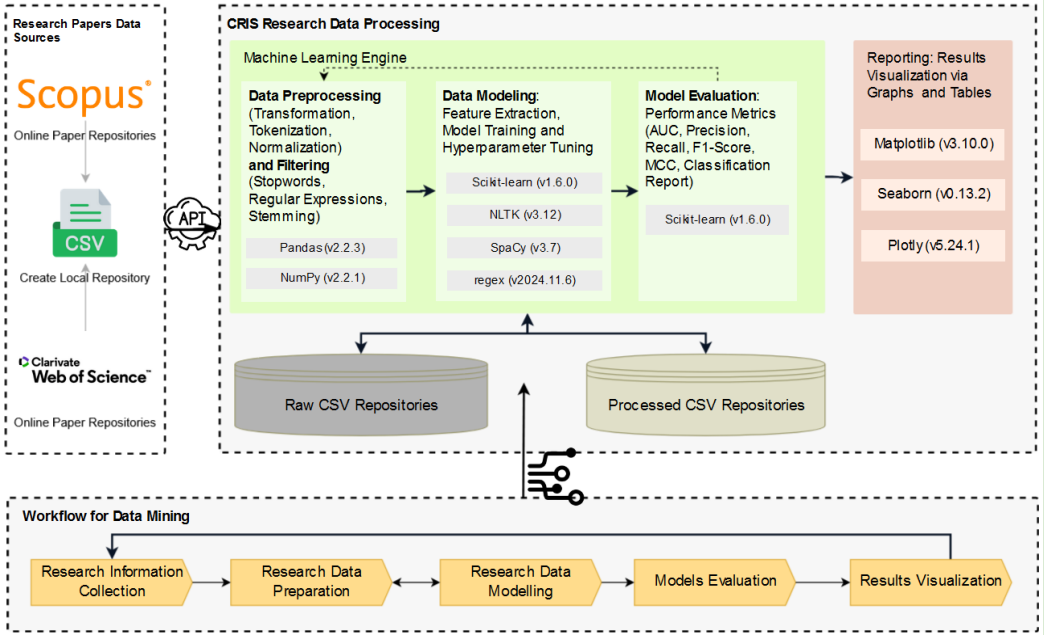

Figure 7: Replicability process for the data mining methodology.

- **Machine learning**: Scikit-learn (v1.6.0), used for classification models such as Support Vector Machines (SVM), k-Nearest Neighbors (kNN), Decision Trees, and Random Forests.
- **Visualization**: Matplotlib (v3.10.0), Seaborn (v0.13.2), and Plotly (v5.24.1).
- **Evaluation**: Classification metrics (e.g., classification_report, confusion_matrix) and performance indicators (e.g., precision, recall, F1-score, MCC).

The data mining workflow is further illustrated in Figure 7, which provides a visual representation of the replicability process. This diagram outlines the step-by-step flow of activities, from data retrieval to results visualization, making the methodology accessible for future adaptation.

6 Results

To test and demonstrate the practical applicability of the PA model described in the previous section, we use data mining software Orange - an open-source machine learning and data visualization tool [67] that covers steps involved in collection, preparation, modeling, evaluation and visualization of retrieved research papers' metadata (Figure 8). In this experiment, we also investigate dependent variables defined in 4.1. before and after applying machine learning algorithms.

Retrieving scientific papers on smart cities was done by using the Scopus abstract and citation database. In that way, we implement **the first stage of our proposed model - Collection of Research Information**. The data is extracted as a structured dataset with metadata type labels as a CSV file.

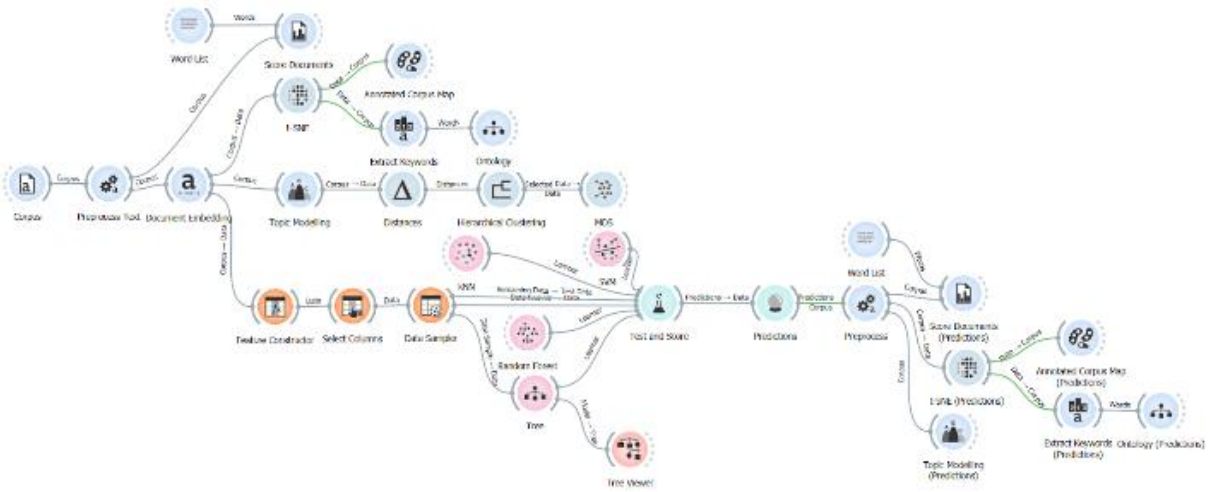

Figure 8: Predictive analytics flow in orange data mining tool.

Document	Word count	Word presence
Towards a cyber-physical system for sustainable and smart building: a use case for optimising water consumption on a SmartCampus	0.286	0.286
A bibliometric evaluation and critical review of the smart city concept – making a case for social equity	2.286	0.286
Optimal management of parking lots as a big data for electric vehicles using internet of things and Long-Short term Memory	0.429	0.286
Quantitative assessment of a smart city on the example of Ekaterinburg	0.714	0.286
Vehicle Artificial Intelligence System Based on Intelligent Image Analysis and 5G Network	0.429	0.286
Exploiting Multi-modal Contextual Sensing for City-bus's Stay Location Characterization: Towards Sub-60 Seconds Accurate Arrival Time Prediction	0.286	0.286
An In-Depth Study of 5G-Based Eco-Friendly Smart City	0.571	0.286
RELATIONSHIP BETWEEN THE DEGREE OF URBAN INTELLIGENCE WITH ARTIFICIAL INTELLIGENCE AND FUZZY ALGORITHMS AND THE PERFORMANCE OF ENTERPRISES	0.571	0.286
FSVM: Federated Support Vector Machines for Smart City	0.429	0.286
A geographic routing based on local traffic density and multi-hop intersections in VANETs for intelligent traffic system in smart cities (GRBLTD-MI)	0.714	0.286

Figure 9: Top 10 scored documents before predictions.

To implement the next stages, we apply several data mining techniques, namely, identifying patterns, classification, association, clustering, as well as data cleaning and visualization.

In the second stage of our proposed model - **Research data preparation** (Figure 6), the dataset was pre-processed, applying transformation, tokenization, normalization and filtering techniques. First of all, accents have been removed, and URLs and all letters have been made lowercase. Then word splitting was done and punctuation was preserved. WordNet Lemmatizer was used for the normalization. To perform filtering, stopwords in English were applied, and numbers and special characters were removed by regular expression.

The research data modelling stage was implemented as feature extraction of the processed dataset. N-grams Range from 1 to 5 was applied to create n-grams from tokens. POS Tagger's type Averaged Perceptron Tagger was run with Matthew Honnibal's averaged perceptron tagger. Keyword and topic extraction techniques are also applied to form an ontology, hierarchical clusters, paper categories, and document scoring.

A word list for scoring the documents was prepared that contains the key phrases used for retrieving the dataset, as it is described in 4.1. - "smart cities", "intelligent cities", "cities IQ", "CRIS development," "CRIS management," and "CRIS systems". Figure 9 shows the resulting list of articles/papers that most often contain keywords/phrases for which the dataset was extracted from the Scopus database. The entire corpus was assessed, including title and abstract. The results show that smart cities and urban intelligence are the focus of the headlines.

The extraction of topics from the processed corpus was done using the Latent Semantic Indexing method, which returns both negative and positive words, as well as topic weights. 10 rows of topics were retrieved and are shown in Figure 10.

1: city, smart, data, system, ©, based, technology, model, urban, network
2: city, smart, data, system, network, traffic, model, based, time, vehicle
3: iot, smart, urban, traffic, city, model, device, application, internet, security
4: data, system, energy, vehicle, big, power, city, traffic, proposed, building
5: energy, traffic, data, vehicle, network, building, urban, smart, city, iot
6: system, model, network, traffic, data, vehicle, method, energy, iot, learning
7: energy, urban, smart, iot, traffic, data, technology, development, city, application
8: traffic, energy, system, urban, model, iot, network, smart, city, vehicle
9: city, system, technology, smart, traffic, iot, vehicle, research, study, urban
10: model, urban, network, iot, traffic, city, smart, vehicle, algorithm, service

Figure 10: Top 10 extracted topics before predictions.

Based on the formed set of topics, hierarchical clusters were formed (Figure 11) by calculating the distances between rows in the dataset using the normalized Euclidean method.

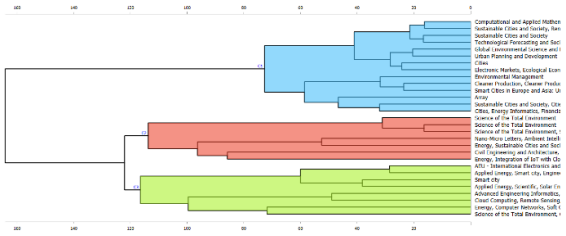

Figure 11: Shows good connectivity of the cluster elements.

A t-distributed stochastic neighbor embedding (t-SNE) method was used to form the categories of scientific papers (annotated corpus map) and ontology. To increase the attractive forces between points, an Exaggeration parameter with a value of 2 out of 4 was used, and to control the number of principal components, the PCA components parameter took a value of 20 out of 50. In this way, we aim to highlight the global structure of the data (see Figure 12).

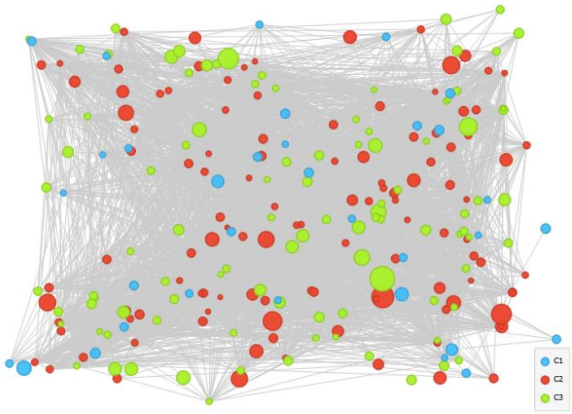

Figure 12: Projection of hierarchical clusters based on topics modelling.

Using the method of Gaussian mixed models, 5 clusters were formed, shown in Figure 13. These are also the main names of the paper categories that are formed from the processed corpus. There is a thematic coincidence with the theme of smart cities and closely related issues.

The t-SNE method was also used to form the ontology of the dataset, firstly for extracted keywords by the YAKE! method, which is suitable for texts of different sizes. The formed ontology can be seen in Figure 14. It is

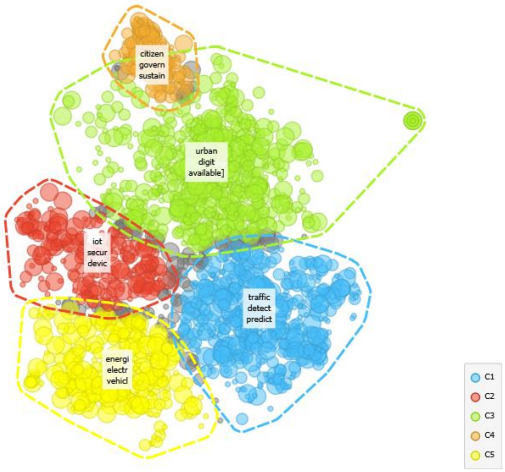

Figure 13: Paper categories based on clustering of datasets before predictions.

also related to the topic of smart cities, urban intelligence and related issues. **At the next stage** of our proposed PA model - **Models Evaluation**, we applied several machines learning algorithms. Table 2 shows the results of their evaluation. The following metrics were used: Area under ROC (AUC aka "Area Under the Curve" of the "Receiver Operating Characteristic" curve); Classification accuracy (CA); F1; Precision and Recall. We applied: SVM (settings: linear with numerical tolerance 0,001, 100 iterations); kNN (Euclidean metric was used with 5 neighbours); Tree (maximum depth 100 instances with minimum number of instances 2) and Random Forest (maximum number of 100 trees). Testing and scoring were performed using a 10-fold cross-validation by feature Type of the publications. 10% of the processed dataset is used for training and the remaining data - for testing.

In addition to the metrics provided in Table 2, confidence intervals were calculated for each metric to provide a better understanding of the variability and reliability of the results. These intervals offer a clearer picture of the model performance and help assess the consistency across different datasets.

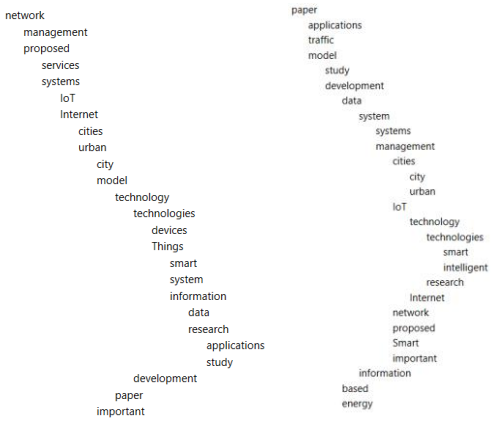

Figure 14: Generated ontology before predictions.

Table 2: Comparison of machine learning algorithms metrics.

Model	AUC	CA	F1	Precision	Recall
Tree	0.233	0.874	0.815	0.764	0.874
SVM	0.248	0.874	0.815	0.764	0.874
Random Forest	0.233	0.874	0.815	0.764	0.874
kNN	0.451	0.874	0.815	0.764	0.874

Furthermore, ROC curves were generated for each model to assess their performance in more detail. The ROC curves (see Figure 15) highlight the trade-off between the true positive rate and false positive rate, giving insight into the overall model performance. The kNN model, in particular, showed the highest AUC value, indicating superior discriminatory ability compared to the others. The experiment is carried out on the independent variables defined in 4.1. to observe the impact that scientific papers are classified by type of publication.

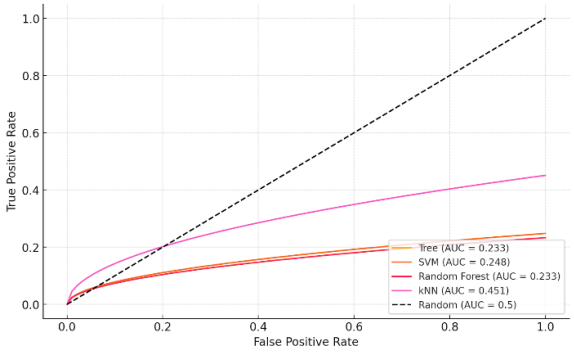

Figure 15: ROC curves visualization, illustrating each model's performance.

The results for all investigated metrics are identical, except for AUC. The kNN algorithm is the most accurate, followed by SVM. The values resulting from Tree and Random Forest are equal. All four algorithms show good results - over 75% precision, which gives us reason to consider that the adequacy results are of a high degree of accuracy.

A tree was also formed from a sample of the dataset, which shows the type of publications according to the area of the publications in which they were published. Numbers above tree instance names indicate the number of citations. As shown in Figure 16, the extracted dataset reveals that a larger portion is formed by articles, followed by reviews. The third most common category is book chapters. The publications in the sample are thematically focused mainly on smart cities and ambient intelligence.

After applying the machine learning algorithms, a new dataset is formed to use for predictions. It undergoes the same text processing as the raw dataset. From the

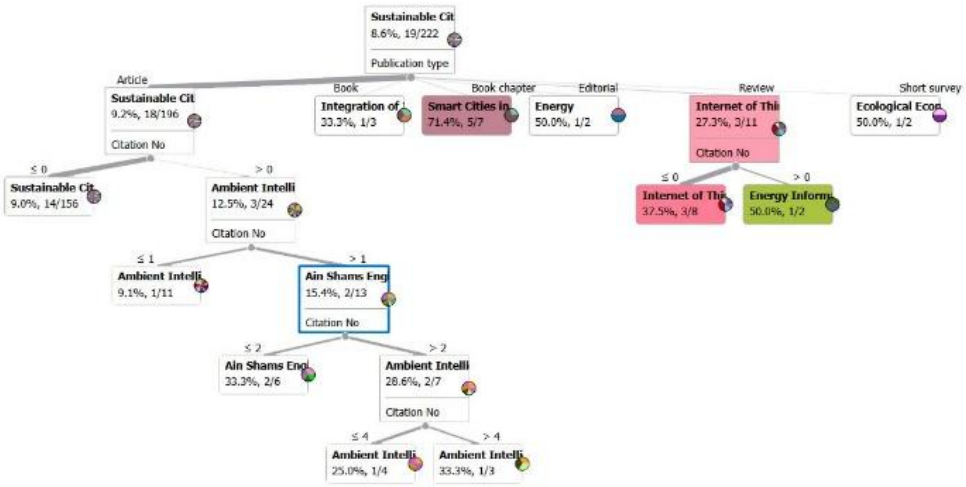

Figure 16: Tree classification sample based on publication type and journal area.

prediction dataset, the research papers are scored, the results of which are provided in Figure 17. The same list of keywords was used for evaluation as before the application of the machine learning algorithms, as it is described in phase #4 of section 4.1. After predictions have been produced by the model, a different picture emerges - the emphasis in the scoring is on CRIS and its application in different areas, but not smart cities anymore.

Again, using the Latent Semantic Indexing method, a prediction was made for the ten most trending topics from the prediction dataset. They are shown in Figure 18.

1: city, smart, data, system, based, ©, technology, urban, model, iot
2: city, smart, data, system, network, traffic, based, proposed, vehicle, model
3: data, iot, urban, smart, system, energy, network, city, model, security
4: data, iot, smart, urban, system, model, energy, wa, study, traffic
5: technology, traffic, utilization, robot, covid, city, network, model, intelligence, health
6: energy, traffic, building, vehicle, city, data, power, parking, system, consumption
7: system, model, network, iot, data, energy, method, traffic, vehicle, parking
8: urban, iot, system, smart, utilization, energy, water, robot, city, wa
9: model, traffic, system, network, energy, water, urban, vehicle, based, smart
10: model, traffic, network, algorithm, iot, system, city, method, energy, smart

Figure 18: Top 10 extracted topics after predictions.

In contrast to document scoring, it is predicted that the topics of greater interest will be still related to smart cities, urban intelligence and their derivative issues.

Next, clusters were formed from the prediction dataset again using the method of Gaussian mixed models (Figure 19). The clusters identified show the categories of research papers that should be of greater interest. A slightly different distribution of the number of papers by

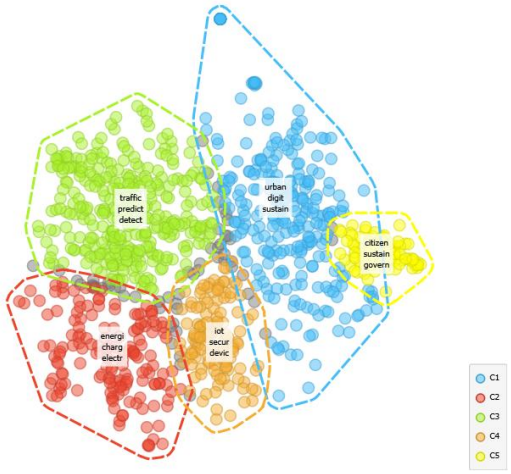

Figure 19: Paper categories based on clustering of datasets after predictions.

cluster can be seen, but the trend is maintained – papers thematically related to smart cities.

On the basis of the prediction dataset, an ontology was generated according to the extracted keywords using the YAKE! method (Figure 20). It is noted that it is thematically linked again to smart cities, urban intelligence and their derivative themes.

Document	Word count	Word presence
Comparative effectiveness of smartphone healthcare applications for improving quality of life in lung cancer patients: study protocol	0.143	0.143
Leveraging Axiomatic Design and Research Information Systems to Promote Research Outcomes at Public Universities	0.286	0.143
Research performance and scholarly communication profile of competitive research funding: the case of Academy of Finland	0.429	0.143
A multicentre, multi-national, double-blind, randomised, active-controlled, parallel-group clinical study to assess the safety and efficacy of PDA10 (Epoetin-alpha) vs. Eprex® in patients with anaemia of chro...	0.143	0.143
Development and characterisation of CRIS systems in Latin America: Preliminary results of diagnostic survey	0.429	0.143
Development and Evaluation of a Mobile Web-based Food Allergy and Anaphylaxis Management Educational Program for Parents of School-aged Children with Food Allergy: A Randomized Controlled Trial	0.143	0.143
Diamond Open Access in Norway 2017–2020	0.143	0.143
Challenges in managing semantic annotations in harvested research objects in a national CRIS context	0.286	0.143
Effects of preoperative education using virtual reality on preoperative anxiety and information desire: a randomized clinical trial	0.143	0.143
The changing scope of data quality and fit for purpose: Evolution and adaption of a CRIS solution	0.429	0.143

Figure 17: Top 10 scored documents after predictions.

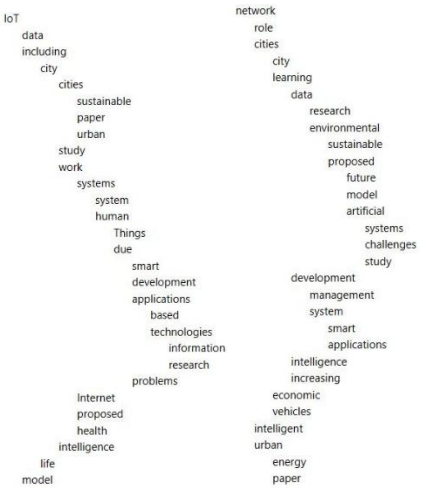

Figure 20: Generated ontology after predictions.

7 Discussion

PA has become an advanced analysis method that utilizes both new and historical data to predict activities, behaviors, and trends. By creating prediction models, research organizations can anticipate various future scenarios, enabling them to make more accurate and timely decisions. For example, in preventing "poor research," research outcomes can be evaluated to identify potential performance issues among researchers, and ideally, the causes can be pinpointed. This allows institutions to take timely action and implement corrective measures.

In the field of research management, PA can assist in detecting patterns and trends in data, and extracting valuable information. The insights gained serve as the foundation for decision-making and control processes within institutions. Executives and other stakeholders can make informed, data-driven decisions. These insights can set the agenda for scientific strategies as well as administrative priorities, such as supporting and funding more promising research topics.

The analysis of research information plays a critical role in efficiently retrieving relevant data and deriving measures or forecasts based on this information. These tasks are usually carried out by CRIS managers and subject matter experts in the field. While deep knowledge of the statistical relationships behind the models is not necessary, research results must be interpretable in a way that allows meaningful conclusions to be drawn. Basic knowledge of PA helps discuss results with analysts and contextualize them appropriately.

Our proposed PA decision-making model aims to help research institutions make impact-related decisions based on a combination of quantitative and qualitative indicators, such as citation counts, journal areas, article types, identification of trending topics, ontologies, and cluster categories of interest to its users.

A key aspect of scientific document classification is organizing collective knowledge within the scientific community, making research more accessible, and fostering collaboration and innovation across disciplines. In this paper, we examine metadata from 20,000 publications indexed in the scientific database Scopus. Titles and abstracts were verified to match the keywords used to extract the dataset before and after making predictions. Machine learning algorithms were applied to create a new dataset from which predictions about trending topics, ontologies, and papers were made.

Before making predictions, the data revealed that topics such as smart cities, urban intelligence, and their derivative areas were leading. Although keywords and phrases related to CRIS were used, they did not receive the highest weighting in Orange Data Mining. This suggests that publications solely focusing on CRIS constitute a small portion of the dataset and cannot serve as a sufficient basis for clustering. This is also evident from the Tree Classification Sample in Figure 15.

After the predictions, the trend of smart cities and urban intelligence topics was disrupted when scoring papers, but the trend was preserved for the extracted topics and ontology from the prediction dataset, confirming the popularity of these topics among the scientific community.

7.1 Ethical considerations

While our model offers significant advantages, it is essential to address the ethical implications of using such predictive models in research institutions. First, there are concerns about data privacy when integrating external sources into CRIS systems [68]. Researchers' personal data, institutional affiliations, and other sensitive information may be at risk if not handled carefully. It is crucial to ensure that data is anonymized and securely stored, and that access is limited to authorized personnel only. Furthermore, it is necessary to obtain consent from data subjects when using their information for prediction purposes, adhering to relevant privacy laws and guidelines.

Another important consideration is bias mitigation in predictive models. Machine learning algorithms, such as those used in our model, can inadvertently perpetuate or amplify existing biases in the data. For example, certain research topics or institutions may be overrepresented or underrepresented in the training dataset, leading to biased predictions that may impact decision-making. To mitigate such bias, it is essential to use diverse and representative datasets, ensure transparency in model development, and regularly audit the models for fairness [69]. Implementing these measures can help ensure that the PA models produce equitable and unbiased outcomes.

7.2 Quantitative comparison with existing studies

To evaluate the effectiveness of our model, we compared the quantitative results with those from the related work summary table (Table 1). For instance, the

study by Piryonesi & El-Diraby showed an improved prediction accuracy for project delays in the construction industry through PA [17]. However, their approach was limited to a narrower domain, whereas our model applies PA within a much broader context of research institutions and CRIS systems. The accurate prediction of research trends and ontologies in our model demonstrates a superior ability to identify research gaps compared to the approaches observed in the related works.

Furthermore, the studies by [18] and [19] in the e-commerce sector demonstrated that PA led to increased sales and better customer engagement. However, our model, which focuses on predicting trending topics and cluster categorizations, not only provides higher accuracy in identifying topics but also enables broader adaptability to the multidimensional nature of research data. This capability of handling heterogeneous data sets our model apart from the existing models.

7.3 Why our model outperforms others

Unlike traditional PA methods, which are often focused on business data (e.g., sales forecasting in retail and e-commerce), our model addresses the complex, dynamic, and multidimensional data structures inherent in research information, typically found in CRIS systems. We combine multiple data sources, such as publication metadata and citation analyses, and offer more precise prediction and analysis of trending topics and research ontologies.

Another significant difference is the flexibility of the model, which can be applied not only to existing research data but also to future research areas that have not yet been captured. This ability to adapt and continuously improve predictions is a key differentiator from the existing models, which often rely on static datasets.

In conclusion, PA enables institutions to manage large volumes of research information efficiently, and our proposed PA decision-making model allows CRIS users—such as universities, research institutions, and libraries—to fully leverage their content to find better information and make informed recommendations. This ultimately leads to improved research strategies and outcomes.

8 Conclusion and future work

Due to the availability of large amounts of data in CRIS, the efficient processing of this research information and the free access to relevant actors, a large number of users can now use and further develop PA methods. The field of information systems has made great strides in the application of advanced statistical modeling techniques in recent years. In summary, PA makes it possible to develop models that can make predictions and assess prediction probabilities. PA can help identify new patterns and behaviors in unfamiliar environments and uncover potential causal relationships. These data relationships can in turn lead to new theoretical models. PA can also become more important in CRIS and offers opportunities, but also during implementation, to create well-founded forecasts to support decision-making. PA can support research management in making processes between the actors involved more accurate, reliable, and cost-effective. Globally active institutions, in particular, which operate in an internationally branched value creation network, have to master great challenges with their needs and capacities. In addition, these are mostly independent organizations that have project and data publication systems that are decoupled from other institutions and do not disclose their data comprehensively.

However, the proposed PA model has certain limitations. For instance, its scalability to larger datasets remains an area of future exploration, especially considering the increasing volume of scientific data in CRIS. Future versions of the model may need to handle datasets beyond the current size, ensuring that its performance remains consistent. Additionally, the model’s adaptability to other institutional systems beyond CRIS, such as universities or other research organizations with different structures and workflows, warrants further research. This would broaden the applicability of the model across various domains.

In terms of computational complexity, a comparison between the different stages of the proposed model and alternative methodologies is needed. Specifically, evaluating the efficiency of data retrieval, preprocessing, and model training could help in optimizing the proposed framework. By doing so, institutions could make more informed decisions regarding resource allocation for research management.

For future work, we aim to explore a range of applications and methods, such as AI approaches and techniques, elastic solutions, and knowledge graphs, to further enhance the accuracy and adaptability of the model. Exploratory research into the integration of AI and machine learning advancements for dynamic data handling within CRIS will provide valuable insights into how the model can adapt in real-time to new research data and emerging trends.

In addition, we envision conducting specific case studies or simulations that validate the proposed PA model's effectiveness in real-world settings, which would allow us to test the model’s practical viability and refine its predictions. This will provide a structured roadmap for potential technological advances and practical implementations that could benefit research institutions globally. We aim to investigate how well-founded decision-making methods can improve the quality of research management, enabling institutions to make better, data-driven choices in the face of evolving research landscapes.

Conflicts of interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] Perrons, R. K., & Jensen, J. W. (2015). Data as an asset: What the oil and gas sector can learn from other industries about "Big Data". Energy Policy, 81, 117–121. https://doi.org/10.1016/j.enpol.2015.02.020

[2] Marr, B. (2018) Here's Why Data Is Not the New Oil https://www.forbes.com/sites/bernardmarr/2018/03/05/heres-why-data-is-not-the-new-oil/?sh=1c70e5133aa9

[3] Nikiforova, A. (2023). HackCodeX Forum Keynote "Data Quality as a prerequisite for you business success: when should I start taking care of it?", https://anastasijanikiforova.com/2023/06/07/hackcodex-forum-keynote-data-quality-as-a-prerequisite-for-you-business-success-when-should-i-start-taking-care-of-it/

[4] Salemink, I., Dufour, S., van der STEEN, M., & Officer, S. P. (2019). Future advanced data collection. In the Conference of European Statisticians, vol. 58, p. 29.

[5] Azeroual, O.; Schöpfel, J.; Pölönen, J. and Nikiforova, A. (2022). Putting FAIR Principles in the Context of Research Information: FAIRness for CRIS and CRIS for FAIRness. In *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KMIS*, pages 63–71. https://doi.org/10.5220/0011548700003335

[6] Bibri, S. E. (2021). Data-driven smart sustainable cities of the future: An evidence synthesis approach to a comprehensive state-of-the-art literature review. *Sustainable Futures*, *3*, 100047. https://doi.org/10.1016/j.sftr.2021.100047

[7] Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of business research*, *70*, 263–286. https://doi.org/10.1016/j.jbusres.2016.08.001

[8] Sarker, I.H. (2021) Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective. SN COMPUT. SCI. 2, 377. https://doi.org/10.1007/s42979-021-00765-8

[9] Ikegwu, A. C., Nweke, H. F., Anikwe, C. V., Alo, U. R., & Okonkwo, O. R. (2022). Big data analytics for data-driven industry: a review of data sources, tools, challenges, solutions, and research directions. *Cluster Computing*, *25*(5), 3343-3387. https://doi.org/10.1007/s10586-022-03568-5

[10] Vu Nguyen Hai, D., & Gaedke, M. (2021, May). Applying Predictive Analytics on Research Information to Enhance Funding Discovery and Strengthen Collaboration in Project Proposals. In *International Conference on Web Engineering* (pp. 490-495). Cham: Springer International Publishing.

[11] Al Sadi, I. M. S. (2021). *Open access analytics with open access repository data: A Multi-level perspective* (Doctoral dissertation, University of Southampton).

[12] Krüger, A. K., & Petersohn, S. (2022). From Research Evaluation to Research Analytics. The digitization of academic performance measurement. *Valuation Studies*, *9*(1), 11-46.

[13] Azeroual, O., Nacheva, R., Nikiforova, A., Störl, U., & Fraisse, A. (2023). Predictive Analytics intelligent decision-making framework and testing it through sentiment analysis on Twitter data. In *Proceedings of the 24th International Conference on Computer Systems and Technologies* (pp. 42-53).

[14] Eisenhardt, K.M.; Zbaracki, M.J. (1992). Strategic Decision Making. *Strategic Management Journa*l, 13,17–37. https://www.jstor.org/stable/2486364

[15] Stylos, N., & Zwiegelaar, J. (2019). Big data as a game changer: how does it shape business intelligence within a tourism and hospitality industry context? In *Big data and innovation in tourism, travel, and hospitality* (pp. 163-181). Springer, Singapore. https://doi.org/10.1007/978-981-13-6339-9_11

[16] Varian, H. R. (2014). Big data: new tricks for econometrics. *Journal of Economic Perspectives*, 28(2), 3-28. https://doi.org/10.1257/jep.28.2.3

[17] Piryonesi, S. M., & El-Diraby, T. E. (2020). Data analytics in asset management: Cost-effective prediction of the pavement condition index. *Journal of Infrastructure Systems*, 26(1), 04019036. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000512

[18] Spencer, S. B. (2015). Privacy and predictive analytics in e-commerce. *49 New England Law Review 101*, 629. https://ssrn.com/abstract=2678381

[19] Akter, S., & Wamba, S. F. (2016). Big data analytics in E-commerce: a systematic review and agenda for future research. *Electronic Markets*, 26(2):173–194. https://doi.org/10.1007/s12525-016-0219-0

[20] Tanlamai, J., Khern-am-nuai, W., & Adulyasak, Y. (2022). Identifying arbitrage opportunities in retail markets using predictive analytics. *Available at SSRN 3764048*. http://dx.doi.org/10.2139/ssrn.3764048

[21] Balbin, P. P. F., Barker, J. C., Leung, C. K., Tran, M., Wall, R. P., & Cuzzocrea, A. (2020). Predictive analytics on open big data for supporting smart transportation services. *Procedia Computer Science*, *176*, 3009–3018. https://doi.org/10.1016/j.procs.2020.09.202

[22] Kelley, K.; Clark, B.; Brown, V.; Sitzia, J. (2003). Good practice in the conduct and reporting of survey research. *International Journal for Quality in Health Care*, 15(3): 261–266. https://doi.org/10.1093/intqhc/mzg031

[23] Maassen, P. A. (1997). Quality in European higher education: Recent trends and their historical roots. European Journal of education, 111–127. https://www.jstor.org/stable/1503543

[24] Hüther, O., & Krücken, G. (2016). Nested organizational fields: Isomorphism and differentiation among European universities. *The University Under Pressure* (*Research in the Sociology of Organizations, Vol. 46*), Emerald Group Publishing Limited, Bingley, pp. 53–83. https://doi.org/10.1108/S0733-558X20160000046003

[25] Jeffery, K., 2012. CRIS in 2020. In: CRIS2012: 11th International Conference on Current Research Information Systems (Prague, June 6–9, 2012). http://dspacecris.eurocris.org/ handle/11366/119

[26] Schöpfel, J.; Azeroual, O. (2021). Current research information systems and institutional repositories: From data ingestion to convergence and merger. Editor(s): David Baker, Lucy Ellis, In Chandos Digital Information Review, Future Directions in Digital Information, Chandos Publishing, pp. 19-37. https://doi.org/10.1016/B978-0-12-822144-0.00002-1.

[27] Clements, A.; Proven, J., 2015. The emerging role of institutional CRIS in facilitating open scholarship. In: LIBER Annual Conference 2015, London, June 25th, 2015. https://dspace-cris.eurocris.org/handle/11366/393

[28] Fraumeni, B. M. (2001). E-commerce: Measurement and measurement issues. *American Economic Review*, 91(2), 318–322. https://www.jstor.org/stable/2677781

[29] Chu, M. K., & Yong, K. O. (2021). Big data analytics for business intelligence in accounting and audit. *Open Journal of Social Sciences*, 9(9), 42–52. https://doi.org/10.4236/jss.2021.99004

[30] Paul, L. R.; Sadath, L.; Madana, A. (2021). Artificial Intelligence in Predictive Analysis of Insurance and Banking. In *Artificial Intelligence* (pp. 31-54). CRC Press.

[31] Morris, K. J., Egan, S. D., Linsangan, J. L., Leung, C. K., Cuzzocrea, A., & Hoi, C. S. (2018). Token-based adaptive time-series prediction by ensembling linear and non-linear estimators: a machine learning approach for predictive analytics on big stock data. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 1486-1491). IEEE. https://doi.org/10.1109/ICMLA.2018.00242

[32] Wadan, R., & Teuteberg, F. (2019). Understanding requirements and benefits of the usage of predictive analytics in management accounting: Results of a qualitative research approach. In *Business Information Systems: 22nd International Conference, BIS 2019, Seville, Spain, June 26–28, 2019, Proceedings, Part I 22* (pp. 100-111). Springer International Publishing.

[33] Souza, J., Leung, C. K., & Cuzzocrea, A. (2020). An innovative big data predictive analytics framework over hybrid big data sources with an application for disease analytics. In Advanced Information Networking and Applications: Proceedings of the 34th International Conference on Advanced Information Networking and Applications (AINA-2020) (pp. 669-680). Springer International Publishing. https://doi.org/10.1007/978-3-030-44041-1_59

[34] Boppana, V. R. (2023). Data Analytics for Predictive Maintenance in Healthcare Equipment. *EPH-International Journal of Business & Management Science*, *9*(2), 26-36. http://dx.doi.org/10.2139/ssrn.5005057

[35] Doleck, T., Lemay, D. J., & Brinton, C. G. (2021). Evaluating the efficiency of social learning networks: Perspectives for harnessing learning analytics to improve discussions. *Computers & Education*, *164*, 104124. https://doi.org/10.1016/j.compedu.2021.104124

[36] Agarwal, P., Tang, J., Narayanan, A. N. L., & Zhuang, J. (2020). Big data and predictive analytics in fire risk using weather data. Risk analysis, 40(7), 1438-1449. https://doi.org/10.3389/fcvm.2021.741667

[37] Thunderbird. (2021). The Power of Predictive Analytics | Thunderbird (asu.edu)

[38] Brunk, J., Stierle, M., Papke, L., Revoredo, K., Matzner, M., & Becker, J. (2021). Cause vs. effect in context-sensitive prediction of business process instances. *Information systems*, 95, 101635. https://doi.org/10.1016/j.is.2020.101635

[39] Di Francescomarino, C., Ghidini, C., Maggi, F. M., & Milani, F. (2018). Predictive process monitoring methods: Which one suits me best? In Business Process Management: 16th International Conference, BPM 2018, Sydney, NSW, Australia, September 9–14, 2018, Proceedings 16 (pp. 462-479). Springer International Publishing. https://doi.org/10.1007/978-3-319-98648-7_27

[40] Kim, SW., Gil, JM. (2019). Research paper classification systems based on TF-IDF and LDA schemes. *Hum. Cent. Comput. Inf. Sci.* **9**, 30. https://doi.org/10.1186/s13673-019-0192-7.

[41] Rathore, A. K., Kar, A. K., & Ilavarasan, P. V. (2017). Social media analytics: Literature review and directions for future research. *Decision Analysis*, *14*(4), 229-249.

[42] Burow, L., Gerards, Y., & Demmer, M. (2017). Effektiv und effizient steuern mit Predictive Analytics. *Controlling & Management Review*, *61*(9), 48–56. https://doi.org/10.1007/s12176-017-0122-3

[43] Eckerson, W. W. (2007). Predictive Analytics: Extending the Value of Your Data Warehousing. TDWI Best Practices Report, Renton.

[44] Zazzaro, G., Mercogliano, P., & Romano, G. (2017). Data Mining for Forecasting fog Events and Comparing Geographical Sites. *IARIA Int. J. Adv. Networks Serv*, *10*, 160-171.

[45] Chapman, P.; Clinton, J.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C.; Wirth, R. (2000). CRISP-

DM 1.0 – Step-by-step data mining guide. SPSS Inc.

[46] Halper, F. (2014). Predictive analytics for business advantage. *TDWI Research*, 1-32.

[47] Chen, C. P.; Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information sciences*, 275, 314-347. https://doi.org/10.1016/j.ins.2014.01.015

[48] Wang, Y., Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological forecasting and social change*, *126*, 3-13. https://doi.org/10.1016/j.techfore.2015.12.019

[49] Grover, V., Chiang, R. H., Liang, T. P., & Zhang, D. (2018). Creating strategic business value from big data analytics: A research framework. *Journal of management information systems*, *35*(2), 388-423. https://doi.org/10.1080/07421222.2018.1451951

[50] Beulen, E., & Dans, M. A. (2023). Data Analytics and Digital Transformation. Taylor & Francis.

[51] Schöpfel, J., Azeroual, O., & Saake, G. (2020). Implementation and user acceptance of research information systems: An empirical survey of German universities and research organisations. *Data Technologies and Applications*, *54*(1), 1-15. https://doi.org/10.1108/DTA-01-2019-0009

[52] Azeroual, O., Nikiforova, A. and Sha, K., 2023, June. Overlooked Aspects of Data Governance: Workflow Framework For Enterprise Data Deduplication. In *2023 International Conference on Intelligent Computing, Communication, Networking and Services (ICCNS)* (pp. 65-73). IEEE.

[53] Frizzo-Barker, J., Chow-White, P. A., Adams, P. R., Mentanko, J., Ha, D., & Green, S. (2020). Blockchain as a disruptive technology for business: A systematic review. *International Journal of Information Management*, *51*, 102029. https://doi.org/10.1016/j.ijinfomgt.2019.10.014

[54] Kotu, V., & Deshpande, B. (2014). *Predictive analytics and data mining: concepts and practice with rapidminer*. Morgan Kaufmann.

[55] Jetten, M.; Simons, E. (2019). Research data management incorporated in a Research Information Management system. A case study on archiving data sets and writing Data Management Plans at Radboud University, the Netherlands.EUNIS19: 25th EUNIS Annual Congress (June 5-7, 2019, NTNU, Trondheim, Norway). http://hdl.handle.net/11366/1015

[56] euroCRIS. (2020). Why does one need a CRIS? The Research Process and how a CRIS can support it. [Online] Available at: https://eurocris.org/why-does-one-need-cris. [Accessed 2 July 2023]

[57] Elsevier. (2023). Why you need a Research Information Management System (RIMS). [Online] Available at: https://www.elsevier.com/research-intelligence/rims-and-cris-systems. [Accessed 2 July 2023]

[58] Romeike, F.; Eicher, A. (2016). Predictive Analytics: Looking into the future. FIRM Yearbook, pp. 168–171.

[59] Llamzon, R. B., Ter Chian Tan, F., & Carter, L. (2021). Toward an information systems alignment framework in the wake of exogenous shocks: Insights from a literature review. International Journal of Information Management, 63, 102450. https://doi.org/10.1016/j.ijinfomgt.2021.102450

[60] Azeroual, O. (2019). Text and Data Quality Mining in CRIS. *Information*, 10(12):374. https://doi.org/10.3390/info10120374.

[61] Nacheva, R. (2022). Emotions Mining Research Framework: Higher Education in the Pandemic Context. In: Terzioğlu, M.K. (eds) Advances in Econometrics, Operational Research, Data Science and Actuarial Studies. Contributions to Economics. Springer, Cham. https://doi.org/10.1007/978-3-030-85254-2_18.

[62] Qiu, F., et al. (2022). Predicting students' performance in e-learning using learning process and behaviour data. Sci Rep 12, 453. https://doi.org/10.1038/s41598-021-03867-8.

[63] Tanvir, Q. (2021). Multi Page Document Classification using Machine Learning and NLP. [Online] Available at: https://towardsdatascience.com/multi-page-document-classification-using-machine-learning-and-nlp-ba6151405c03. [Accessed 2 August 2023]

[64] Rahimi, N., Eassa, F., Elrefaei, L. (2020). An Ensemble Machine Learning Technique for Functional Requirement Classification. *Symmetry*, *12*, 1601. https://doi.org/10.3390/sym12101601.

[65] Baadel, S., Thabtah, F., Lu, J., & Harguem, S. (2022). OMCOKE: a machine learning outlier-based overlapping clustering technique for multi-Label data analysis. Informatica, 46(4). https://doi.org/10.31449/inf.v46i4.3476

[66] Kozarovytska, P. & Kucherenko, T. (2023). Empirical comparison of hyperparameteroptimization methods for neural networks. Proceedings of Master's Symposium on Advances in Data Mining, Machine Learning, and Computer Vision (MS-AMLV 2023), pp. 1-11.

[67] University of Ljubljana. (2023). Orange Data Mining. [Online] Available at:https://orangedatamining.com. [Accessed 2 October 2023]

[68] Zendulková, D., & Azeroual, O. (2022). Legal aspects and data protection in relation to the CRIS system. *Procedia Computer Science*, *211*, 17-27. https://doi.org/10.1016/j.procs.2022.10.172

[69] Azeroual, O., Schöpfel, J., Pölönen, J., & Nikiforova, A. (2023). FAIRification of CRIS: A Review. In *International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management* (pp. 280-298). Springer, Cham. https://doi.org/10.1007/978-3-031-43471-6_13