Future Generation Computer Systems 143 (2023) 248–264

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

# AIMDP: An Artificial Intelligence Modern Data Platform. Use case for Spanish national health service data silo

Alberto S. Ortega-Calvo [b], Roberto Morcillo-Jimenez [b], Carlos Fernandez-Basso [a,b,*], Karel Gutiérrez-Batista [b], Maria-Amparo Vila [b], Maria J. Martin-Bautista [b]

[a] Causal Cognition Lab, Division of Psychology and Language Sciences, University College London, London, United Kingdom
[b] Department of Computer Science and Artificial Intelligence, University of Granada, 18071, Granada, Spain

ARTICLE INFO

*Article history:*
Received 14 July 2022
Received in revised form 30 January 2023
Accepted 2 February 2023
Available online 6 February 2023

*Keywords:*
Data platform
Artificial Intelligence
Intelligent data analysis
Big data
Medical informatics
Data silo

ABSTRACT

The huge amount of data being handled today in any environment, such as energy, economics or healthcare, makes data management systems key to extracting information, analysing and creating more efficient daily processes in these environments. However, the inability of current systems to take advantage of the data generated can waste good opportunities for analysing and extracting information from the data. Modern data platforms (MDP) appear suitable for supporting management systems and are able to perform future prospective analyses. This paper presents a data platform called Artificial Intelligence Modern Data Platform (AIMDP), based on Big Data, artificial intelligence for management, and efficient data handling. The different components of AIMDP intervene in the data acquisition phase and implement algorithms capable of analysing massive data collected from heterogeneous sources. In addition, the entire platform is geared towards data management and exploitation with a layer of security and data governance that allows the integrity and privacy of the databases to be maintained. The proposed platform is designed to be used by users who are not experts in data science. To this end, it implements a user-oriented workflow that has effectively been introduced in a use case of two Spanish hospitals to extract knowledge from their historical data, which had been siloed and had never been explored by any hospital researchers or doctors.



## 1. Introduction

Nowadays, companies and researchers are increasingly interested in developing robust systems that efficiently enable data storage and knowledge discovery. Modern data platforms (MDP) are computer-based systems which enable organizations to become data-driven [1].

These systems provide developers and end-users with tools for storing, processing and analysing complex data from heterogeneous sources and domains. Modern data platforms have been widely used in different domains such as smart cities, financial technology, materials science, energy, and many others [1–6].

Building these sorts of systems raises many challenges, such as processing and analysing heterogeneous data sources, providing end-users (mostly non-expert users) with a user-friendly tool, treatment of massive data and data governance. These challenges become even more complex when the data to be analysed is electronic health records (EHRs). It is mainly because medical databases can contain data from different sources, which results in the data being mostly semi-structured and unstructured. Medical databases store historical data about patients. We can find very few studies in the literature related to developing MDP oriented to healthcare [7–11]. We summarize the main drawbacks to address for implementing these kinds of systems in the healthcare domain below:

- EHRs are often stored in data silos, where data comes from different sources, and the data are mostly comprised of semi-structured and unstructured data.
- Providing end-users (mostly users without knowledge about data science) with user-friendly services and tools, enabling end-users to focus only on the desired analysis.
- The enormous amount of data collected in medical data silos makes processing and analysing difficult.
- Establishing appropriate data governance policies in order to provide better availability, integrity, usability and security of the data.

Considering the problems above, it would be useful to endow healthcare-related centres and users with powerful tools to manage EHRs.

* Corresponding author at: Department of Computer Science and Artificial Intelligence, University of Granada, 18071, Granada, Spain.
*E-mail addresses:* cjferba@decsai.ugr.es, carlos.basso@ucl.ac.uk (C. Fernandez-Basso).

https://doi.org/10.1016/j.future.2023.02.002
0167-739X/

Our study aims to develop a multi-purpose MDP-based architecture that enables professionals in many areas of knowledge who are not experts in data science to handle large amounts of data flexibly and efficiently. The tool would allow end-users to perform more detailed analyses of their data. To this end, this study proposes a new architecture called an Artificial Intelligence Modern Data Platform (AIMDP). The new architecture integrates the main features of the MDP with Data Science (DS) capabilities. We must remark that our research group has already implemented the algorithms comprised in the platform [12–17]. These capabilities are based on supervised and unsupervised techniques in a distributed environment (Big Data).

Although AIMDP is able to work with data of different natures, to showcase the feasibility of the proposed data platform, we have conducted a real-world use case using two healthcare data sets from two hospitals of the Spanish national health system service. Since AIMDP allows end-users to process and analyse health-related data, it could facilitate possible diagnostics and analysis, the subsequent treatment, and the premature prevention of potential diseases. Following on, we summarize the main contributions of this paper:

- **Heterogeneous data sources** - The proposed architecture can handle semi-structured and unstructured data from different sources. This feature and a dynamic database structure allow the end-users to analyse data from different research areas without worrying about the structure and provenance of the data.
- **User-friendly** - AIMDP provides end-users with an interactive and intuitive interface. In this way, the tool can be readily used by users who do not have any expertise in data mining (non-expert users).
- **Bigdata-based techniques** - AIMDP enables data processing and analysing through Data Science capabilities. All the algorithms have been implemented using the distributed programming paradigm.
- **Data governance** - The availability, accessibility, integrity, usability and security of the data in the platform are carried out through an innovative security layer that allows the users of the platform to work with sensitive data.

The rest of the paper is structured as follows: a review of previous research related to this topic is presented in Section 2. Sections 3 and 4 presents the proposed data platform (AIMDP) for artificial intelligence applications and how users can use the data platform workflow modules, respectively. Section 5 presents a real-world use case using AIMDP and discusses the obtained results. Finally, in Section 6, the conclusions and future research are presented.

## 2. Related research

The *platform concept* has been well known since the early 2010s when there were already discussions [18] about the different connotations of this term. The connotation used in this paper is the computational one, which complements the *data platform* term, one of the crucial concepts of this study. AIMDP, the platform that is being introduced in this paper, is a data platform that adopts a Modern Data Architecture schema.

In the related research section, an introduction to the data platform concept is made. The most relevant terms related to this concept, such as *Cloud Computing*, *Big Data* or *Modern Data Architecture*, are described from the perspective of the data platform concept. AIMDP is also detailed within these relevant terms. In addition, some applications of data platforms are mentioned, introducing real examples of previous proposals made by other authors in this field. The application in health care is explained in detail, as there is a lot of potential in this area, which is the main one on which the use case of Section 5 is based. The capabilities of the recently developed data platforms are discussed separately and compared to those of AIMDP.

The objective of this section is for the reader to be able to firstly understand the concept of data platform and its context and secondly know the main aspects of the proposed AIMDP data platform before the technical description made in Sections 3 and 4. And finally, appreciate the main differences between AIMDP and other data platforms that appear in the scientific literature.

### 2.1. Data platform

Nowadays, data storing and treatment operations are experiencing a shift [19] from data warehouses to data lakes. This shift is produced by three main factors: 1. The increase in the amount of semi-structured and unstructured data, which are data warehouses unfriendly. 2. The popularity of the micro-service architecture over the monolithic one, which does not have an associated central database. 3. The inability to fulfil the 5 Vs (variety, volume, velocity, veracity and value) requirements.

Regarding the data warehousing and data lake concepts, it is possible to combine these two technologies, even in a cloud-based environment, under the name of Modern Data Platform [19]. This is more capable than a data centre, as it provides more features that address the needs of the new data consumers. Some of these features are, amongst others, that in a data platform, there is no need to provide a schema for the incoming data or the use of Spark [20], and will offer more flexibility since it will be able to deal with large and semi-structured data sets and use multiple parallel tasks for processing [21].

In this paper, the data platform concept is considered one of the main aspects. The standardization of this term is specially promoted by the commercial sector. This is because data platforms as products are a current trend. However, in recent years, a significant number of data platforms have also been developed in the scientific literature (OPEN GOVERNMENT DATA [2], Financial BDP [3], CiDAP [4], CALIBER [7,8], WikiHealth [9], ENTSO-E [22,23], DEGS [24], D2D BIG DATA [25], TELCO [26]).

In the scientific literature, there are different definitions and interpretations of what a data platform is. For [27], a data platform is a web-based interface to collect, store, host and manage datasets with specific uniform organizations. A data platform is capable of storing a large amount of data in an organized structure. Users of the data platform are allowed to contribute, modify, query, analyse and export the datasets, which is crucial for accelerating the evaluation of building performance and energy efficiency. Similarly, in [10], 4 different data platforms are analysed and compared. The authors set the 3 dimensions of data control as the most important parameters to analyse in a data platform: data access (who can access the data), data use (primary and secondary uses of data, who decides these uses, and how they are legitimized and approved through data governance strategies), and data governance (the process by which stewardship responsibilities are conceptualized and carried out).

In the book [19], a data platform is capable of accomplishing all the operations established by [27], which are needed to implement data analysis. For [19,27], the data analysis process is not included as a mandatory feature of data platforms. In other words, the purpose of a data platform is to ingest, store, process, and make data available for analysis. These operations should be independent of the type of data and be performed in a cost-efficient manner. However, other authors that propose implementations of their own data platforms consider this procedure part of a data platform [3,4,9,26]. Moreover, the authors in [3,5,9,19,26] include a layer-based layout, where the data platform is composed of

modules that work independently and share information through their connections, building a workflow. AIMDP also implements a layer-based architecture with independent modules, described in detail in Section 3.

MongoDB [28] includes a definition of a data platform that comprehends most of the ideas introduced by the previously referenced authors. The definition that they propose is: "a data platform is an integrated set of technologies that collectively meet an organization's end-to-end data needs. It enables the acquisition, storage, preparation, delivery, and governance of your data, as well as a security layer for users and applications".

### 2.2. Heterogeneous data integration

As it is discussed in [29], in the current article, structured data is considered as a group of concepts that possesses a set of attributes and relationships with other concepts in the set or source. This is the source type included in standard relational databases. Semi-structured data has some similarities, but each instance of a concept may have different properties or relationships. Examples of this type of data are XML or JSON documents. Unstructured data does not have a structured representation of its concepts, properties, or relationships. Most common examples of these sources are text files or multimedia content such as images [30], audio, or video.

Working with heterogeneous data and its integration constitutes a complicated and well-known problem. In this article, two main kinds of integration are distinguished: 1. Transforming available data into data with similar properties and structure. 2. Allowing the coexistence of different types of data on the platform.

If data scientists and data architects consider following the first approach, it will result in a simpler database system and a possible unification of the data processing techniques of the platform. As included in [29], progress has recently been made to transform unstructured data into semi-structured and structured data. Giving this unstructured data a 'structure' can be achieved through techniques that consist of extracting keywords (from sources such as text, videos, or images) and establishing a representation schema and mathematical models. This can simplify the architecture of the data system and enable simultaneous query on heterogeneous data and intereschema property extraction (synonymies, homonymies, hyponymies, overlappings...), thus enriching the data and associated ontologies. However, this requires quality assurance in the metadata and a data pre-processing stage that is not trivial, especially working in a Big Data scenario.

In this platform, data integration is based on the second approach, the coexistence of heterogeneous data sources. Having different types of data available on the platform may require more effort in different aspects, such as database management or algorithm implementation. However, AIMDP grants the possibility of carrying out data processing, knowledge extraction and analysis of results based on the techniques described in Section 4. The implementation is done so that the user can work with unstructured, semi-structured and structured data and will always have available advanced techniques to perform the required experiments.

### 2.3. Cloud computing in data platforms

Although MongoDB [28] explicitly establishes that this is not a mandatory technology in a data platform, the Cloud Computing paradigm improves its quality. Some of the features [19] that this paradigm introduces in data platforms are elasticity (they grow and shrink as the needs of the data platform change), modularity (custom resources in storage and computing), availability (resources are available at any time) and faster development (faster introduction of new features in the production environment) of the resources. Being able to use third-party Infrastructure as a Service (IaaS) and Cloud Computing services in a data platform is a great advantage since it allows scalability and flexibility that cannot be guaranteed with a standard server infrastructure. The data platform proposals in [9,25] use third-party services on demand, allowing data platform administrators to increase the storage and compute capabilities of their systems based on their traffic and needs.

In the proposed data platform, regarding the NIST definition of Cloud Computing [31], AIMDP implements a solution for each of the characteristics that define a Cloud Computing infrastructure. It adopts a Software as a Service (SaaS) model since the tools of the data platform are accessible from various client devices through a web-based user interface and the consumer does not manage the underlying cloud infrastructure (network, servers, operating systems, storage, etc.). AIMDP can also be defined as a community cloud because its infrastructure is provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns. A more precise description of how AIMDP's cloud components are configured is included in Section 3.3.

### 2.4. Big data platform

The amount of data has increased dramatically in recent decades as society has become more involved with technology, and companies have discovered that producing and storing data can generate considerable profit. Big Data has applications in healthcare, electronics, biology, banking, meteorology and many others fields that affect people's daily lives. Implementing Big Data architectures for organizations requires obtaining expensive software licenses, preparing a large and sophisticated infrastructure, and paying for experts who know how to use the system and organize and integrate the generated data for analytics [32]. Data platform technologies can be used to help developers to meet these challenges. This is where the term Big Data platform (BDP) comes in, which is capable of providing the same services and features as data platforms but working with massive data sets. Moreover, it is worth mentioning that using the data science tools that computational frameworks such as Apache Spark offer can lead the user to perform advanced knowledge extraction and create meaningful visualizations from these big-sized datasets. These computational frameworks are often based on the concept of distributed learning [33], which establishes that training data is located and processed in different nodes. Those nodes are distributed autonomous computers, and they communicate among themselves over a network. Many relevant algorithms, such as those introduced in [34] or Apache Spark MLLib are based on distributed learning concept. Some of the platforms mentioned previously [4,9,25,26] are defined as Big Data platforms, and they include big data tools like Spark, Hadoop Distributed File System, Hive, OpenStack, etc.

Big Data platform services and tools are usually combined so the user can use them without worrying about what is going on behind the scenes in terms of availability, security or performance. AIMDP, the data platform described in this article, has some of the characteristics of big data platforms, as is described in Section 3.

### 2.5. Modern data architecture

As mentioned previously, some authors describe a layer-based architecture as the core of their data platforms. In this article, we

include the paradigm Modern Data Architecture (MDA), defined by MongoDB in [28]. It consists of a software architecture that helps to ensure that the user does not have to worry about what is taking place in the core of the system, since it specifies how the data platform and data analysis technologies are structured and carried out.

The AIMDP data platform follows a similar structure to the MDA. This makes the system capable of implementing data analysis and processing techniques using the components of the analytics layer. One of the Big Data characteristics of the data platform is that it implements powerful and big-size-data-proof ML/artificial intelligence(AI) techniques. Something similar appears in [7], where the authors describe the CALIBER phenotyping approach and use it to produce 51 phenotyping algorithms that take part of the CALIBER data platform [8]. This platform includes data resources and tools, algorithms, and specialized infrastructure and supports access to anonymized United Kingdom National Health Service (NHS) data under licence from the Clinical Practice Research Datalink (CPRD) [35]. WikiHealth [9] and TELCO [26] also include a differentiated data analysis layer.

On the other hand, due to the generalist and scalable nature of the data platform proposed in this article, it may have been used in different fields of knowledge. AIMDP allows users to load data with different structures since it works with NO-Sql database schemas and implements generic Apache Spark ML algorithms. Including algorithms developed in recent years is also straightforward because the AI module is independent of the rest of the modules of the data platform. This feature is described in more detail in Section 3.

### 2.6. Potential in health care

Especially after the Covid-19 pandemic outbreak, doctors and other researchers in the medical field have become crucial since it has been proven once again that their research work has helped to save many lives and managed government actions and decisions [36]. The authors in [37,38] show how multi-disciplinarity, involving different subjects of study in one activity (medical and data analysis in this case), can produce real outcomes in people's health.

Another problem related to health data and, especially, big health data is the existence of data silos [10]. These storages contain data collected from health institutions, biomedical and genomics research and so on, the use of which is limited. This means that all the research potential found in the data has been lost. In order to perform knowledge extraction from these data sets, data platforms such as AIMDP are very useful since they allow any user with permission to provide and mine all the available data in the platform.

Multi-disciplinarity and the data silo problem were two of the main reasons that motivated the development of the AIMDP data platform and carry out the use case described in Section 5. Although the system can work with data of a different nature [39], the use case included in this article concerns medical data.

For more than a decade, the researchers of the group have been working with hospitals of the Spanish national health service. Working with doctors introduces the necessity of building a data platform which is user-friendly, and easy-to-use [40]. The objective becomes to be used correctly by users with basic knowledge of data science and programming, so they can still exploit all the features and extract the maximum potential of the data platform [41,42]. This reveals one of the main features of the system, accessibility.

### 2.7. Data platforms comparison

The characteristics of AIMDP are described in detail in Section 3. The data platform shares points of view with some of the data platforms mentioned in this section but also includes some innovative aspects. A comparison of the most relevant characteristics and capabilities of the data platforms mentioned in this section can be found in Table 1.

The aspects selected for the comparison are the following: 1. **Heterogeneous data collection**: data of different structures (stream, structured, unstructured), data variety, and several data sets. 2. **Data governance**: Availability, integrity, usability and security of data. 3. **Big Data**: Use of Big Data technologies and big-sized data sets. 4. **AI tools**: Use of AI technologies and algorithms for prediction, regression, recommendation, etc. 5. **System security**: Elements and techniques that guarantee the security of the data platform as a whole. 6. **Multi-purpose**: Scalability and flexibility of the system and if it can be used for different purposes with data of diverse nature and fields of knowledge. 7. **User friendly**: Usability of the system and an easy-to-use user interface for accessing the features of the system. 8. **3rd parties IaaS support**: Support of 3rd parties IaaS such as Google Cloud, Azure, AWS, etc. 9. **Tested with real users**: Features of the system used by different users with real data.

Considering the results of Table 1, AIMDP is one the most complete data platform amongst the selected ones. Data governance, heterogeneous data collection and security are characteristics that most data platforms fulfil, as they are central aspects of the definition of the concept. However, features such as AI or big data tools are not supported on all platforms. AIMDP not only includes these features but implements an advanced computing infrastructure that allows the running of Big Data-friendly algorithms. It is compatible with open-source tools such as Spark and allows newer algorithms to be added without much effort, leveraging the infrastructure defined in the following section and producing remarkable results in terms of quality and efficiency.

The main difference between AIMDP and other data platforms is that it is multi-purpose and easy to use. Thanks to the ease of use, the dynamism and the ability to implement generic algorithms, the platform can be classified as multi-application, that is, it can be used to build different applications in many areas such as energy, medicine, social networks and so on, without being associated with any particular field.

As can be seen in the comparison table, AIMDP presents a limitation in the use third-party IaaS, whilst CALIBER and PatientsLikeMe does not. This decision has been taken because the data platform is designed to work with data with several levels of privacy. To guarantee data security, a set of private and centralized servers are used. However, in future challenges, described in Section 6, the development of a hybrid system is contemplated. This system allows critical data to be managed on a private server or a set of decentralized servers (allowing the use of federated learning) and the rest of the data on third-party servers.

Additional limitations of the platform are related to: first, the import of external data since, although it is dynamic, the management of this information may be improved by the direct connection to the ontology-based enrichment system; second, the lack of an AutoML tool so that the user can query the system and it automatically generates the entire pipeline. These two limitations are planned to be addressed in a future (see Section 6.1).

In Section 5, the features from Table 1 are put to the test in a real use case.

**Table 1**
Comparison of the main features of the mentioned data platforms. Legend: ✓- Feature supported, ✗- Feature not supported, ∼ - Feature not fully supported or with explicit limitations, ? - Unknown, information not available about the feature.

| Name | Heterogeneous data collection | Data governance | Big data | AI tools | System security | Multi-purpose | User friendly | 3rd party IaaS support | Tested with real users |
|---|---|---|---|---|---|---|---|---|---|
| AIMDP | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ |
| ENTSO-E [22] | ✗ | ✓ | ✗ | ✗ | ? | ✗ | ∼ | ✗ | ✓ |
| CALIBER [7] | ✓ | ✓ | ∼ | ✓ | ✓ | ✗ | ? | ✗ | ✓ |
| D2D Big Data [25] | ✓ | ∼ | ✓ | ✓ | ∼ | ✗ | ✗ | ✓ | ✗ |
| WikiHealth [9] | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | ✗ |
| CiDAP [4] | ✓ | ∼ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ |
| TELCO [26] | ✓ | ∼ | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ |
| PatientsLikeMe [10,11] | ✓ | ✓ | ✗ | ✗ | ✓ | ✗ | ✓ | ? | ✓ |

## 3. AIMDP architecture design

As explained in Section 2.3, the concept of a data platform has become widespread in recent years. Many data processing companies have oriented their developments towards the implementation of such platforms to optimize the work of their users. One of the main limitations of most platforms is the difficulty in adapting any kind of problem to this type of platform. Due to the complexity of some databases, mainly in the health sector, it is too tedious to achieve the assembly of this type of data structure within the architecture of the platform.

In our study, we have eliminated this weak point, creating a data platform that achieves complete separation of each of its layers, grouping its functionalities into different modules. In this way, the platform is capable of adapting to any type of problem, regardless of its scope. AIMDP is capable to recognize the category of data, whether it is structured, semi-structured or unstructured and to perform a basic adjustment so data from different sources can be stored in the platform and become available for experimentation. This functionality is achieved through the implementation of the Acquisition layer and the Application layer, the functioning of which are explained in detail in Sections 3.2 and 3.7. This provides versatility when performing any type of work, whether it is oriented towards health, as is our use case in Section 5, or other types of work, such as those oriented towards energy efficiency, etc.

In our data platform, we have achieved that the different layers that it is made up of being independent. Thanks to this independence, in the event of an error in any of the layers, the platform is not completely affected, thus considerably reducing critical errors. The independence of each layer has made it possible for the platform to adapt to the different changes that occur throughout the life cycle of the platform in real-time without affecting the rest. Fig. 1 shows the different layers that make up our data platform. The following sub-sections explain the architecture and functionality of each layer.

### 3.1. Source layer

In the source layer of AIMDP, the data platform is enriched with data extracted through a series of extraction, transformation and loading (ETL) processes and stored so that it is available for the users of the platform. To perform this task, we are using Pandas [43]. Pandas is a highly advantageous open-source python library for data science. It is renowned for its speed, power, and flexibility, as well as its ease of use in data analysis. These strengths make it a valuable tool for data scientists and analysts, who can quickly and effectively manipulate and analyse complex datasets. Additionally, its open-source nature means it is constantly being improved and updated by a community of dedicated developers, making it a reliable and cutting-edge option for data analysis tasks. Overall, Pandas is a highly effective and widely-used python library for data science.

In our platform, it is possible to include sources such as data repositories, as well as data lakes, which store a large amount of data that is retained until it is needed to be used. The data stored in these repositories is unstructured. Another storage system we use in our data platform is data warehousing, which is designed to store data in a structured way, as there are also less flexible databases which follow a relational structure. For this task, we use Oracle [44] tools, although any available data warehouse could be used. In this way, we manage to work with both structured and unstructured data.

In this layer, it is also possible to find simpler sources from which a large amount of information is obtained. These are the databases themselves, the import of files, physical media, as well as the use of web services for data collection. With the help of all the above elements, plus the collection of data from sources such as hospitals or sensors in office buildings, the complete data storage ecosystem of our architecture is established.

One of the main advantages that our platform offers over others is that the data is not obtained from a single data source. AIMDP provides the possibility of obtaining data from almost any repository (Github [45], Data Repositories, etc...), sensor (temperature, heating, cool, etc...) or storage device (USB, hard disk, etc...), even if the data is siloed and has strict privacy requirements. Through a series of data extraction, transformation and loading (ETL) operations, AIMDP is capable of adapting to the data set so that it can be stored to be subsequently used by any layer of the platform. The data platform configures the database by identifying the data structure of each source, managing to adapt nearly any data structure regardless of the complexity it offers.

In the following sub-section, we explain how to obtain the data of the different elements explained in this section and how to access the information already stored in the data platform.

### 3.2. Acquisition layer

The data acquisition layer focuses on the creation of the data storage architecture. This is necessary so that the platform can obtain the different data that the user needs to perform different experiments.

This layer is where the necessary ecosystem is created to supply the application layer with the different databases. It also implements a series of application programming interfaces (API) capable of retrieving and accessing stored information that can be interpreted and subsequently processed efficiently. The API has been created using the flask api-restful tool [46] because it offers several advantages. Its implementation in Python makes it easy to use and integrate with other Python-based technologies. Additionally, the flask API-restful tool is well-known for its simplicity and flexibility, allowing developers to quickly and easily build robust, scalable API solutions. Overall, the decision to use

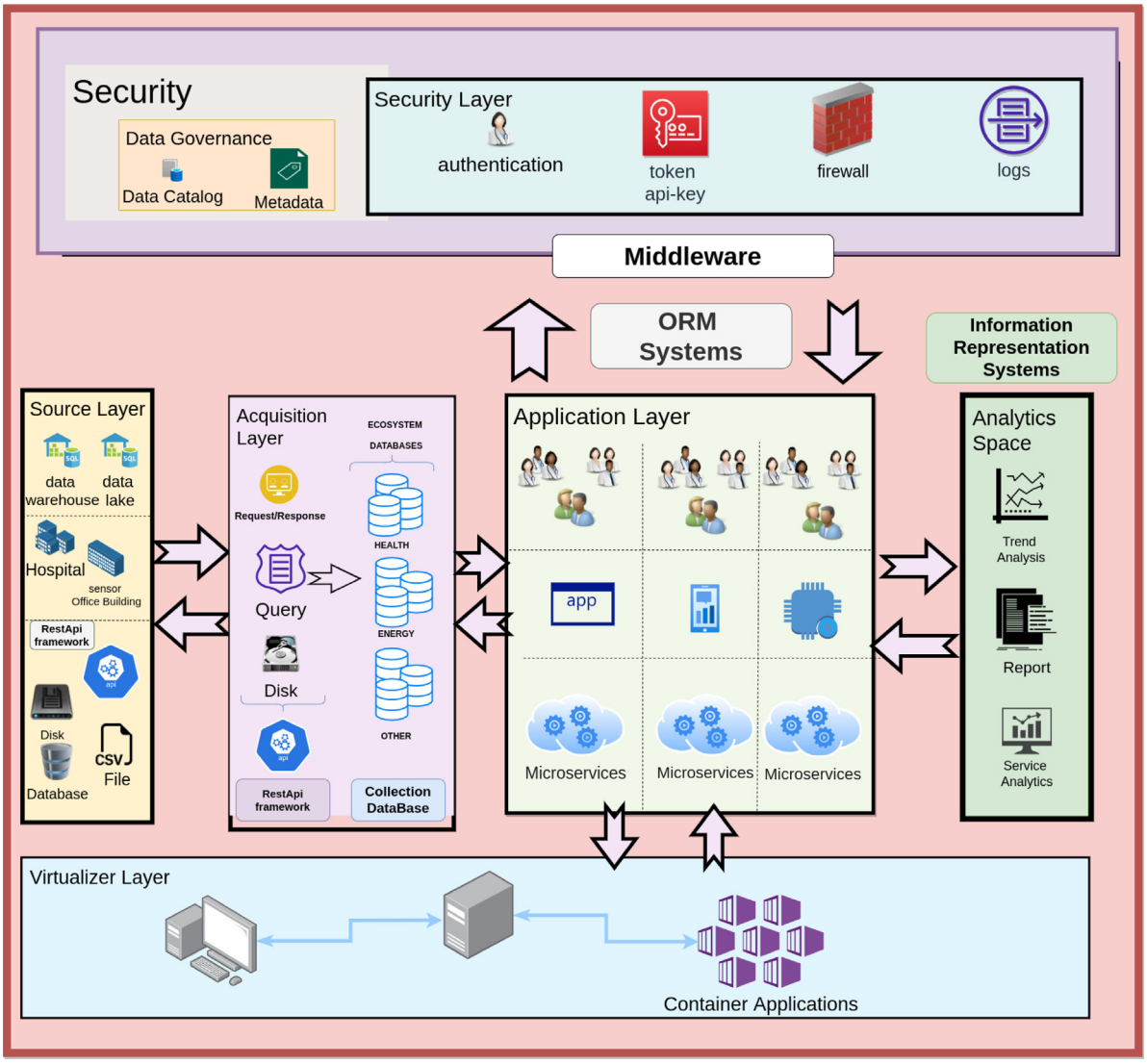



**Fig. 1.** AIMDP architecture design.

flask API-restful was driven by its strong support for Python and its powerful features for building APIs.

In this paper, we have created a storage architecture with the help of MongoDB [47], in order to establish an ecosystem where all the databases are placed and managed to work on the different stored data structures. Each database that makes up this architecture contains a series of collections.

These collections are usually of two types, one where the raw or processed data is stored in key–value documents and another collection where the metadata of the variables of the first collection is stored. Metadata collection is very important because it provides the necessary information for the user to use the data in a structured way in the platform. Through an intelligent process, AIMDP obtains the knowledge of the metadata collection and makes it available to the users so that they can carry out different experiments having all the necessary information about the data.

One of the advantages of this study that differentiates it from others is the possibility of having each of the structures independently within the same database storage ecosystem. This is achieved through the creation of a sub-layer where data is stored in multiple collections, maintaining data and metadata. Thus, when required by the user, the platform loads metadata into the memory for the user to work with. This produces better maintainability and the possibility of producing independent experiments using the data hosted within our storage architecture, achieving more efficient and faster results. Regarding this, the data stored does not necessarily have to be of the same type since the databases are independent. This allows data and structures from different research areas to be part of AIMDP and processed with its tools. Additionally, a feature that is offered to the users is that the data from different areas of knowledge can coexist, being able to assign each group of platform users a subset of databases independently, depending on the permission of the user.

Thanks to the creation of this layer and the storage architecture, one of the objectives we set is fulfilled, which is that users can focus on their experimentation without spending a lot of time adapting, transforming and manually exploring the data stored on the platform.

### 3.3. Data virtualisation layer

The virtualisation layer is based on a cloud architecture that some data platforms implement. We have set up the platform with the help of Docker [48]. This tool automates the deployment of applications inside software containers. This provides an additional layer of abstraction and the automation of application

virtualisation across multiple operating systems. In this way, we achieve a modularized platform.

Docker-driven modularization guarantees that any of the modules, which are independent of each other, can be maintained at any time without affecting the rest of the containers. This is a great advantage because it allows the system to be more stable and respond effectively when critical failures occur, guaranteeing the integrity and security of the information in the modules that are not affected by the failure. Due to this modular approach, it is possible to introduce any application considered useful for the platform. Another aspect to highlight is the possibility of integrating this layer with widespread services such as AWS [49], Google [50] and Azure [51], as it is mentioned in Section 6.1. With this, it would be possible to access the powerful cloud computing resources of these third-party IaaS, thus increasing the potential of AIMDP in terms of the virtualization layer and computing and storage capabilities.

Due to the modularization of this layer and the integration of Docker, the maintainability of the different work ecosystems that make up the platform is achieved. AIMDP provides the possibility of extending its size just by adding any programming interface that contributes new knowledge to the platform.

### 3.4. Analytics space

The analytical space layer is responsible for the generation of different elements to improve the visualization and understanding of the data stored in the data platform. These elements can be simple, such as PDF/HTML reports, automatic charts, or tools that allow users to customize how they view and explore the data they are using. However, AIMDP also supports more complex display options such as graphs, nodes, or data cubes [16]. In this layer, there are a number of micro-services that communicate with the API of the platform, which analyses the data used by the user within AIMDP. This layer performs visualization tasks, reports the results of the different experiments, explores data, OLAP cubes [52], etc.

One of the key features of the analytical space layer is its ability to generate complex visualizations of data. This allows users to gain a deeper understanding of the data stored in the platform and to explore it in new and meaningful ways.

In addition to generating charts and graphs, the layer also supports the use of OLAP cubes, which are specialized data structures that allow users to quickly and easily perform complex data analysis tasks. By leveraging these powerful tools, users can quickly and easily gain insights into their data that would be difficult or impossible to uncover using other methods.

The purpose of this AIMDP layer is to offer more flexibility to the user to understand, visualize and analyse the data. This layer solves the problem of exploring data from a wide variety of sources, a common drawback, as mentioned in Section 2.7. Thus, the elements of the layer are better adapted to the needs of the users' project, allowing a personalized visualization of the data available to them.

In conclusion, the analytical space layer of AIMDP provides users with flexible tools for understanding, visualizing, and analysing their data. This allows them to overcome the limitations of traditional data platforms and to gain a deeper understanding of the data they are working with, as mentioned in Section 2.7. By providing a personalized approach to data visualization and analysis, the analytical space layer is better suited to the needs of users' projects and can help them to extract valuable insights from their data.

### 3.5. Data governance

The main function of the data governance layer is to ensure the integrity of data, the elimination of leaks in data transmission and the availability of data for registered users. Alongside the security layer, it guarantees that only authorized users are able to access each of the databases.

AIMDP's data catalogue follows a hierarchy that goes from the most generic option, where the nature of the project is included, to the multiple databases offered by each of the projects of the data platform.

Policies have been established using the middleware authentication provided by the Django framework used in our system. Our middleware ensures and maintains the integrity of the data within the system's database by performing automated checks on a regular basis to verify the integrity of the data and eliminate any errors within it.

The middleware authentication provided by Django is an essential part of the security protocols in place for our system. It ensures that only authorized users are able to access the data stored in the database and that the data is kept safe and secure. In addition to providing authentication, the middleware also performs regular checks on the data to ensure its integrity. This helps to ensure that the data remains accurate and reliable, even as it is being accessed and used by different users within the system. Overall, the middleware authentication provided by Django is an important tool for maintaining the security and integrity of the data within our system.

In this way, users can only access the parts of the ecosystem they are authorized to, eliminating inconsistencies within the database itself. This is a key feature of the platform, as parts of the data within a single database are restricted to users without permission, ensuring controlled, personalized, and secure access.

### 3.6. Security layer

One of the objectives of the layer is to reduce the loss of information and critical integrity errors that may exist in the different databases, preventing them from affecting the rest of the database. This layer is crucial in the functioning of the platform since sensitive data can be included in AIMDP, as shown in Section 5.

In addition, the elements of the layer guarantee great flexibility to access the databases of the platform. We have divided the layer into three levels, network, application and user or research group:

- At the network level, AIMDP uses a virtual private network (VPN). This allows users and administrators to communicate securely and access platform resources without loss of information.
- Regarding the application level, security is based on Docker containers [48]. We have created an internal network where applications using AIMDP data can communicate when running within the same environment. This is another way to prevent sensitive information from leaving the control of the platform.
- At the user or research group level, an authentication system has been implemented. This system generates internal tokens exclusively for each session in which the user performs an experiment or uses any of the AIMDP tools. This provides the AIMDP administrators with a way to differentiate each user or research group and set custom permissions.

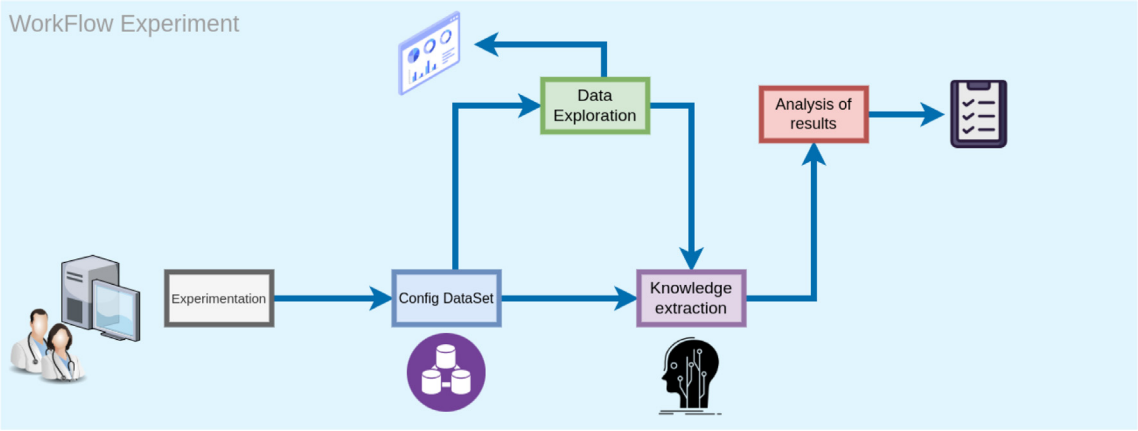



**Fig. 2.** AIMDP experimentation workflow.

### 3.7. Application space

This layer controls the communication of the different layers that make up AIMDP with the applications that use the platform services.

Regarding the acquisition data layer, we obtain the necessary data for its operation through a series of application interface programs. Through the API, different users can interact independently with the storage architecture of the platform, using a series of micro-services that have been created. Micro-services are small programs that execute specific tasks within the data platform. The small size of these tasks makes the micro-services independent of each other, being able to coexist on the platform without depending on each other.

This layer provides the possibility to connect with the analytics space layer, where the information provided to the different applications can be represented in a more customizable way, as explained in Section 3.4. Security and data governance layers, described in detail in Sections 3.5 and 3.6, allow this layer to interact with this information. They offer the possibility of allowing access to certain knowledge of the platform by assigning a series of credentials to each of the platform users, either individually or through work groups.

The virtualisation layer also plays an important role, as explained in Section 3.3. This is where the different applications that make up this layer are deployed on the server. Each of the applications that compose AIMDP is independently assigned a space within the platform server so that the crash or removal of one application does not affect the rest.

Finally, due to the interaction between the different layers, an ecosystem is achieved in such a way that data stored in AIMDP can be used through user-friendly interfaces, following a workflow that allows users to focus exclusively on extracting information from their data. The user interface has been created using Django, a popular framework for developing web applications in python. While other technologies could also be used for this purpose, Django was chosen for its powerful features and strong support for python. However, it is important to note that the developer is not limited exclusively to Django and could potentially use other technologies as well.

This user-friendly paradigm that allows the user to obtain results quickly and efficiently is one of the main aspects of the following section, which offers a deeper study of how the workflow has been implemented.

## 4. AIMDP user-friendly workflow

One of the problems that most non-programming-expert researchers have when performing experiments using a data platform is the lack of a guided workflow that takes them to step by step to achieve a goal in the execution of their work. They have a great deal of expert knowledge about their data, but most of them are not experts in pre-processing and extracting all the knowledge from the data through the application of data science and computational techniques. One of the objectives of this study is to solve this problem. To do this, we follow the paradigm of ease of use, as explained in Section 2.6.

In this section, we explain how the AIMDP workflow is structured to help the user focus exclusively on extracting insights from their data. This section focuses on the application layer, as it allows the user to perform different experiments in a simple way within our AIMDP architecture. As Fig. 2 shows, the workflow is divided into different stages or modules. In the next sub-section, each stage of the workflow is explained, exposing each of its characteristics.

### 4.1. Experimentation module

The experimentation module is the first stage of the workflow. In addition to the application layer, security, governance and acquisition layers are involved in the operation of this module.

The acquisition layer allows the display of the different experiments in this module which are available to the user. With the help of the security layer, it restricts access to each experiment and, as explained in Sections 3.5 and 3.6, prevents data loss. The governance layer is also involved in this module as it achieves data integrity, avoids redundant or erroneous data within the databases, and controls data availability.

The creation of the experiments is dynamically stored in the acquisition layer, associating it exclusively to the owner user through the security layer. Our application layer stores only the metadata, such as the experiment's name, the database selected in our experiment and the parameters chosen to filter the different variables of the experiment itself.

The great advantage of this storage system is that the only thing stored in the database supported by the AIMDP system is plain text, which considerably reduces our application's necessary percentage of storage. This way of storing experimentation does not involve the storage of records from the hospital's data collections, which eliminates data redundancy through the governance

layer. All the management within our application layer is done with the python framework Django.

In this module, users can create a new experiment by selecting a specific database or uploading an experiment already performed by the active user. This gives the user the ability to reuse the work of other experiments, which can be useful for re-running experiments with some changes or not having to enter all filters and workflow parameters.

### 4.2. Data-set configuration module

In this module, the user can configure which variables from the database selected in the previous module are going to be used in the experiment, using a variable selection tool. It is also possible to filter the variables, taking into account whether they are numeric or categorical. In the case of numeric variables a numerical range can be selected, and in the case of categorical variables, some particular values within the domain of the variable can be selected.

One of the main problems of data processing systems is the great variety of databases, each of which has different variables, data types and even representation structures. AIMDP solves this problem by dynamically and intelligently assembling the variables that make up the database on which the experimentation is based, thanks to the storage architecture described in Section 3.2. Therefore, regardless of the user's specialization, AIMDP can automatically upload the data to the data platform.

With the AIMDP filtering tool, the user can select both the data set and the variables needed for experimentation. The filters and parameters selected in this module are stored in a configuration file, which prevents the user from having to configure these filters multiple times. In this way, the platform gains in efficiency and we prevent the server from suffering load saturation, reducing the percentage of critical errors on the platform.

At this point, users can perform exploratory data analysis on the data they have selected or direct their experimentation towards knowledge extraction.

### 4.3. Knowledge extraction module

One of the main barriers of other data platforms is that extracting knowledge from the data is often quite tedious and inflexible due to the numerous constraints presented by these algorithms, as well as the need for extensive knowledge in the field of data science and programming. The elimination of this drawback is one of the goals set in the development of AIMDP, which guides the user through extracting knowledge from the selected dataset for the current experiment.

The user can select any of the algorithms offered by AIMDP implemented[2] in our APIRest. For the execution of this kind of algorithm, advanced knowledge in data science is required. Therefore, a user help system has been implemented to show the user the necessary help to run this algorithm and to see if this type of algorithm fits the data used in our experimentation.

Currently, our AIMDP allows us to execute cluster algorithms in a sequential or parallelized way due to the first adaptation offered by our tool in distributed environments implemented in the Spark framework. Once selected, each of these algorithms can configure automatically executed by a Big-Data-proof computational cluster, following the conditions required by our help system for the execution of the algorithm in question.

Another advantage of our AIMDP architecture is its high encapsulation. This feature allows us to increase the range of algorithms needed to adapt to different experiments without interfering with developing other application layers that interact with our architecture.

The main goal of this module is that the user can focus on the configuration of his dataset and does not have to spend a large part of his time delving into the field of data science. Once the algorithm has been run, the user can check and analyse the results using the results analysis module, detailed in the results 4.5.

### 4.4. Exploration module

One of the reasons why this type of platform is indispensable is the possibility of representing data in a way that makes it easy for the user to visualize and obtain information from it. In AIMDP, a module includes this type of interpretation, which is the data exploration module. In this module, users can comprehensively analyse and visualize the dataset selected for experimentation. Users can select any data or variable from the dataset and analyse it in a way that allows them to conclude their studies.

This module allows the user to visualize their variables' behaviour in the advanced stages of running the experimentation. It allows the user to extend their knowledge of the selected variables, discovering hidden or undetected behaviours and allowing them to decide to improve the execution of the different algorithms.

The analytics space layer supports this module. This layer achieves the objective of encapsulating the functionality within AIMDP in order to be able to extend its behaviour throughout the architecture life cycle without affecting different applications that are interacting with AIMDP. Some of the more straightforward visualization tools are offered by the analytics space layer in AIMDP.

Currently it allows the user to generate pie charts and histograms, depending on the ranges and types of variables selected for exploration. It also includes more complex visualization tools to see the possible relationships between two or more variables, using box plots, scatter plots and others, depending on the type of variables to be compared (see Table 2).

As mentioned above, AIMDP also allows the user to visualize more complex data using graphing tools, nodes or data cubes, giving the possibility to develop external visualization libraries and interact with our architecture [16].

### 4.5. Result analysis module

The results analysis module is usually fundamental in all data platforms. It is where users can draw their conclusions from the experiments performed and decide whether the execution of an experiment achieves the expected results or whether a new, different experiment should be performed.

This module analyses the results obtained after running an algorithm from the knowledge extraction module. It generates a series of automatic tables and graphs depending on the algorithm selected by the users and the data used in the experiment. Depending on the algorithm executed, the system will generate a series of visualization elements to be displayed to the user. These dynamic and intelligent charts and graphs help the user interpret the experiment's results, which is crucial in the development of their studies.

The application layer manages this stage of the workflow. In this way, independence is achieved when displaying the results to the user, as the application layer decides how to display these results. This decision is because, depending on the problem, the different application layers using our architecture will be free to display the results in one way or another.

Our AIMDP will only manage the processing of the data and will send results that the application layer will be in charge of representing. As the application layer manages these results,

**Table 2**
Algorithms included in AIMDP.

| Name | Cite | Kind of algorithm | Aim | source | Big data | Efficiency |
|---|---|---|---|---|---|---|
| BDARE | [53] | Association Rules | No supervised | Own development | ✓(Spark) | Exponential |
| FIM in Streaming | [12] | Frequent items set mining | No supervised | Own development | ✓ | Exponential |
| FARE | [14] | Fuzzy association Rules | No supervised | Own development | ✓(Spark) | Exponential |
| OBTD | [15] | Topic detection | No supervised | Own development | ✗ | Cubic |
| CDMA | [16] | Multidimensional analysis | No supervised | Own development | ✗ | Cubic |
| FSD | [17] | Fuzzy multidimensional analysis | No supervised | Own development | ✗ | Quadratic |
| MLIB LR | [54] | Logistic regression | supervised | MLIB | ✓ | Linear |
| MLIB DT | [54] | Decision tree | supervised | MLIB | ✓ | Linear |
| MLIB RF | [54] | Random Fore | supervised | MLIB | ✓ | Logarithmic |
| MLIB SVM | [54] | Support Vector machine | supervised | MLIB | ✓ | Cubic |
| MLIB GBT | [54] | Gradient-Boosted Trees | supervised | MLIB | ✓ | Logarithmic |
| XGBOOST | [55] | XGBOOST | supervised | Xgboost | ✓ | O(tdxlogn) |
| scikit-learn PCA | [56] | PCA | No supervised | scikit-learn | ✗ | Cubic |
| scikit-learn Kmeans | [56] | K-means | No supervised | scikit-learn | ✓ | Quadratic |
| Distributed K-Prototypes | [57] | Clustering | No supervised | Implementation based on [57] | ✗ | Linear |

users can store each of the results provided by this module in its reserved storage area. In our case, it can be consulted at any time for future studies. The possibility of reusing previous reports generated after the execution of an experiment is another advantage of AIMDP.

In the next section of this paper, we include an experiment that goes through all the workflow stages described in this section. To test this workflow and the architecture described in Section 3, we have included a real use case with health-oriented data silos.

## 5. Use case: IA experiment applied to a spanish health service data silo using AIMDP

In fields such as medicine, a great deal of data tends to remain siloed and not used in data analysis tasks due to privacy and standardization problems [10]. Data silos [10,58] are repositories that cannot be accessed from the outside because of privacy, regulations and other issues. These data sets with unexploited potential may cause losses in financial or medical organizations [59], due to delays in scientific progress in different areas. The problem of data silos motivated the use case introduced in this paper. Since the main collaborators and users of the data platform are doctors and researchers from Spanish hospitals, the data silos in their storage system can be analysed and processed. The AIMDP platform can also provide an improvement in the research work of the collaborators of the hospital since they do not have to expend hours improving their programming skills and they can focus on their real study area. This is achieved with the tools described in Section 3.7, which allow users to perform data mining and analysis tasks without the need for programming skills since usability and ease of use are the central aspects of the AIMDP data platform.

In order to validate and evaluate the AIMDP data platform, the architecture and all the modules presented in Section 3, a real use case has been carried out. It consists of setting up all the features of AIMDP so it can be used by real medical collaborators of two hospitals in the Spanish health service: the Costa del Sol and San Cecilio hospitals. These researchers are the first real users of the system. In order to test all the features the following tasks have been carried out: 1. Two health data sets that have never been used for data analysis (data silos) with medical information of patients in these two hospitals have been uploaded to the data platform. 2. The system has been deployed on a server of the University of Granada. 3. An account for each of the users has been created. 4. A pre-processing and enrichment of data has been carried out. 5. An experimentation workflow has been defined, based on the aspects described in Section 4, so the users can perform a whole experiment using all the modules of the data platform. 6. An experiment using the data set has been done and the results have been analysed.

### 5.1. Data sources

The data sets uploaded into the data platform used for the experimentation follow the structure of the Minimum Basic Data Set (MBDS) of Andalusia [60]. The structure of these data sets is based on EHR and they contain the health records of real patients of the two hospitals. These records come from different sections of the hospitals (emergency, mental health, maternity, etc.) and each record is associated with a patient's hospitalization episode. The data stored in these registers have different formats because they come from a wide diversity of data sources. Some of the most representative sources and features can be seen in Table 3.

Characteristics about the dimension of the raw data sets can be found in Table 4. Since AIMDP implements features of a data lake, raw data is uploaded and kept in the storage architecture described in the acquisition layer (Section 3.2). This is useful since the data platform implements Big Data tools and it is capable of working with large amounts of data and takes advantage of information that could be lost during the data processing stage. Nevertheless, during the data loading process, some custom operations are carried out automatically on the data sets. This generates new data sets that are also stored in the platform. These operations are described in the following sub-section.

### 5.2. Data processing

As has been mentioned previously, the data sets considered for the use case described in this paper have the structure of the Minimum Basic Data Set (MBDS) of Andalusia. With this in mind, a series of operations are carried out on data at the same time the data is being uploaded into the AIMDP storage architecture, details of which can be checked in 3.2. The operations are transformation, pre-processing, enrichment and loading. All these operations are hidden from the user, maintaining the usability of the system.

#### 5.2.1. Data transformation and pre-processing

During the transformation stage, factor variables are mapped from integers to understandable labels, using the manuals and supporting documents of MBDS [60]. During this phase, the data is transformed into key–value documents. This is an efficient way to store the considered data sets because there are very dispersed variables.

The next stage is the pre-processing of data. Here, new variables are computed using existing variables, since the information represented in these variables is usually not data analysis-friendly. Variables used for these operations are history codes, postcodes, dates, etc. Some examples of this kind of pre-processing can be found in Table 5.

**Table 3**
Data source features description.

| Features | Codification | Method for enrichment |
|---|---|---|
| Diagnoses | International Classification of Diseases (ICD) | External API |
| Origin of patients | Spanish health service system | Database |
| Reason for discharge | International Classification of Diseases (ICD) | External API |
| Reason for admission | International Classification of Diseases (ICD) | External API |
| Surgical procedures | International Classification of Procedures | External API |
| Diagnostic tests | International Classification of Test | External API |
| Services/departments | Spanish health service system | Database |
| Other data[a] | Andalusian health service system (SAS) | Database |

[a]Rest of the variables related to the management and information of the Andalusian hospital system.

**Table 4**
Characteristics of the data sets of the use case.

| Data set | Records | Features |
|---|---|---|
| EHR - Costa del Sol Hospital | 75.000 | 273 |
| EHR - San Cecilio Hospital | 220.000 | 273 |

#### 5.2.2. Data enrichment and loading

Data enrichment techniques applied to the data sets use the classifications and codes included in the International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10) [61]. This process is carried out on variables such as diagnoses, surgical procedures or external causes of morbidity and mortality. A mapping with 3 different levels of granularity is made with these variables, as is shown in Table 6. Dig level 1 provides a more wide and more generic diagnosis, while level 3 maps the diagnosis to a more specific and precise one.

After all the mentioned processing has been applied to each data set, they are uploaded into the storage system of the data platform. This process is described in Sections 3.1 and 3.2. For each data set, a copy of the raw data transformed into key–value documents and a processed one are uploaded.

### 5.3. Data experimentation

In this section, a complete clustering experiment is presented using the AIMDP data platform. This experiment has been shown to the users of the data platform as an example or guide of how an experiment should be performed, so it is meant to be understandable and show the main characteristics of an experimentation workflow. The proposed clustering problem consists of extracting information about the underlying structure of childbirths in the Costa del Sol Hospital between the years 2016 and 2020. A secondary objective is the preparation of the data for future data analysis using complex AI algorithms such as [62–66].

Before the configuration of the data set and algorithms of the experiment, the user must access the URL of the system using a web browser. This web application is deployed on a University of Granada server. The user then has to log in using the authentication module, described in Section 3.6. This is also a key piece in the data governance management of the platform, defined in Section 3.5. This is because once users have been logged in, it is possible to control which data is available for each user, what they can do with the data and so on. The user is now capable of creating and running an experiment following the steps included in Fig. 2. All the modules that appear in the figure are described in detail in Section 4.

The first step is the creation of an experiment and the selection of the parent data set on which the experiment is based. This can be done using the experimentation module, which is defined in Section 4.1. The parent data set contains the processed and enriched data of the hospital, obtained after applying the methods described in the previous Section 5.2 to the data set with raw data. This processed data set is selected because these types of algorithms give a worse performance using raw data. The second data set of Table 1 is selected as the parent data set, assuming the user has access to it. Next, the user can configure a sub-dataset using the variable selection and filtering tool of the system, which does not require any programming skills from the user. Available metadata information is also useful since it ensures that the user possesses all the needed information about each variable. These operations are carried out in a very efficient way since, as it is mentioned in Section 3.7, the system can perform the filtering and selection operations by only using the metadata of variables.

These filtering and selection operations are carried out by the data set module, described in Section 4.2. The selected variables and filters for the clustering problem-solving can be found in Table 7. It is worth mentioning that only childbirth episodes are considered since PESO1N, the variable that contains information about the weight of the babies is selected and missing values are removed from the experiment. After the sub-data set has been built, all the filters and selected variables are stored in a configuration file that can be loaded for future experimentation, as it is described in Sections 4.1 and 4.2.

At this point, the user can perform an EDA using the tools of the module detailed in Section 4.4 or proceed with the experiment and use the tools available in the knowledge extraction module of Section 4.3, where the clustering algorithms can be configured. The EDA module has been used to explore the data and purpose of the clustering problem, so the following step is to select the algorithm and set up its parameters. For solving a clustering problem, there are some available algorithms in Spark MLLib. From this library, algorithms such as K-means or Bisecting K-means are supported currently in AIMDP, but these do not take advantage of the categorical variables' information. To solve this problem, a Spark-based version of the classic K-Prototypes has been developed, based on the proposal [57].

**Table 5**
Example of the pre-processing of temporary data from the hospital database.

| Date admission | Date discharge | | Birthday |
|---|---|---|---|
| 20/4/2016 10:16:40 | 23/4/2016 18:23:12 | | 11/2/1991 |
| 10/10/2019 22:45:20 | 13/10/2019 19:20:45 | | 7/2/1987 |
| | ↓ | | |
| Days of admission to labour | Season | Age | Patient type |
| 2 | spring | 31 | adult |
| 3 | autumn | 35 | adult |

**Table 6**
Example of the processing of an instance of C00.4 diagnosis.

| Diagnosis |
|---|
| K35.20 |
| ↓ |

| Dig Level1 | Dig Level2 | Dig Level3 | Application to | Group diagnosis |
|---|---|---|---|---|
| Diseases of the digestive system | Diseases of appendix | Acute appendicitis | (Acute) appendicitis with generalized peritonitis NOS | Major gastrointestinal disorders and peritoneal infections with and without cc/mcc |

**Table 7**
Selected variables for the clustering experiment.

| MBDS code | Description | Type | Filtered domain |
|---|---|---|---|
| PAISNAC | Mother's country of Birth | Categorical | Argentina, Spain, Morocco, Paraguay, United Kingdom |
| EDAD | Mother's age in years | Numerical | All available data |
| TIEMPOING | Hospital admission time during childbirth in days | Numerical | All available data |
| PESO1N | Weight of the baby in kg | Numerical | All available data |
| TGESTAC | Gestation time in weeks | Numerical | All available data |

Once the K-Prototypes algorithm has been selected, an automatically generated HTML form is shown to the user. This form has the parameters needed for the execution of the algorithm, and each form attribute is populated with default values. The HTML form and the values chosen for the parameters of the K-Prototypes algorithm's execution can be found in Fig. 5. After all the parameters have been set, the user can run the algorithm by pressing a button that sends all the information to the system. The connection to the computer cluster and the generation of the result plots and tables is also automatic, so the user completes the whole experimentation workflow without seeing a single line of code.

### 5.4. Results and discussion

Since the solved problem is a clustering one, the results are very graphic and easy to understand. This is the aim of the experiment: to be understandable for non-experts and to bring closer the AIMDP data platform and its capabilities.

The data platform offers a series of tables and plots generated automatically, depending on the parameters and the results provided by the algorithm, in this case, the K-Prototypes clustering algorithm. The most relevant table, which includes the main interpretation of results, contains the labels of each cluster found by the algorithm, the number of individuals placed on each cluster of data and a summary of the value of each variable in the cluster. The mean and the mode are computed for numerical and categorical variables, respectively. This interpretation can be found in Table 8. From this table, it is possible to deduce some information and propose some hypotheses:

- The algorithm found 4 clusters of different sizes. Clusters 3 and 4 are the most populated groups.
- The most common countries of birth in the first two clusters are UK and Morocco, respectively, whilst the most common in the last two clusters is Spain. The algorithm made a clear separation in the groups based on this variable.
- There is no significant difference regarding the mean weight of the baby, gestation time and hospital admission time variables (PESO1N, TGESTAC, TIEMPOING).
- Looking at the mother's age variable, the mean in the first two clusters is similar, whilst the mean in the third and fourth one is clearly different, with a difference of approximately 10 years.
- The weight of the babies is higher in cluster 2.
- Bearing all of this knowledge extracted by the algorithm in mind, the following hypothesis is proposed if 4 clusters are considered: 1. The first cluster contains mothers from the UK, which is the smallest group found. The weight of their babies is similar to the Spanish mothers' babies from clusters 3 and 4. The age is similar to the one of cluster 2 and the mean between clusters 3 and 4. 2. The second cluster contains mothers from Morocco whose babies' weight is higher and triplicates the size of the first clusters in terms of individuals. 3. The third and fourth clusters aggregates Spanish mothers of different ages, with cluster number 3 being associated with younger mothers. Looking at the sizes of the last two clusters, it is possible to observe that there are more mothers that belong to the cluster with a higher value of age.

These results can also be observed graphically in the plots generated automatically by the system. In Fig. 4, which contains a plot of the distribution of the considered numerical variables, it is possible to see the importance that the algorithm gives to the mother's age variable (EDAD), which is very important in the delimitation of clusters 3 and 4. On the other hand, in Fig. 3, the relationship between only the categorical and the rest of the numerical variables is summarized in a box plot. This provides information about the distribution of the rest of the selected countries of origin, Argentina and Paraguay, which are distributed amongst the four clusters, with the minority being in selected data. Looking at the plot, the importance of age is also visible, as it has different distributions amongst the clusters. As the last aspect to remark, the user is also able to see that the weight of the baby and the gestation time are variables with a considerable number of outliers, which must be considered in future analysis operations using this data set.

This experimentation process has been put on into a video, which can be accessed by all the registered users of the platform, serving as an example and a user guide. The problem that has been approached and the results of the algorithm were presented to the health-specialist collaborators of the research team. Since all the hospital personnel were capable of understanding and checking the potential of AIMDP data platform, we see this as a new way of validation of the experimentation module and the system as a whole.

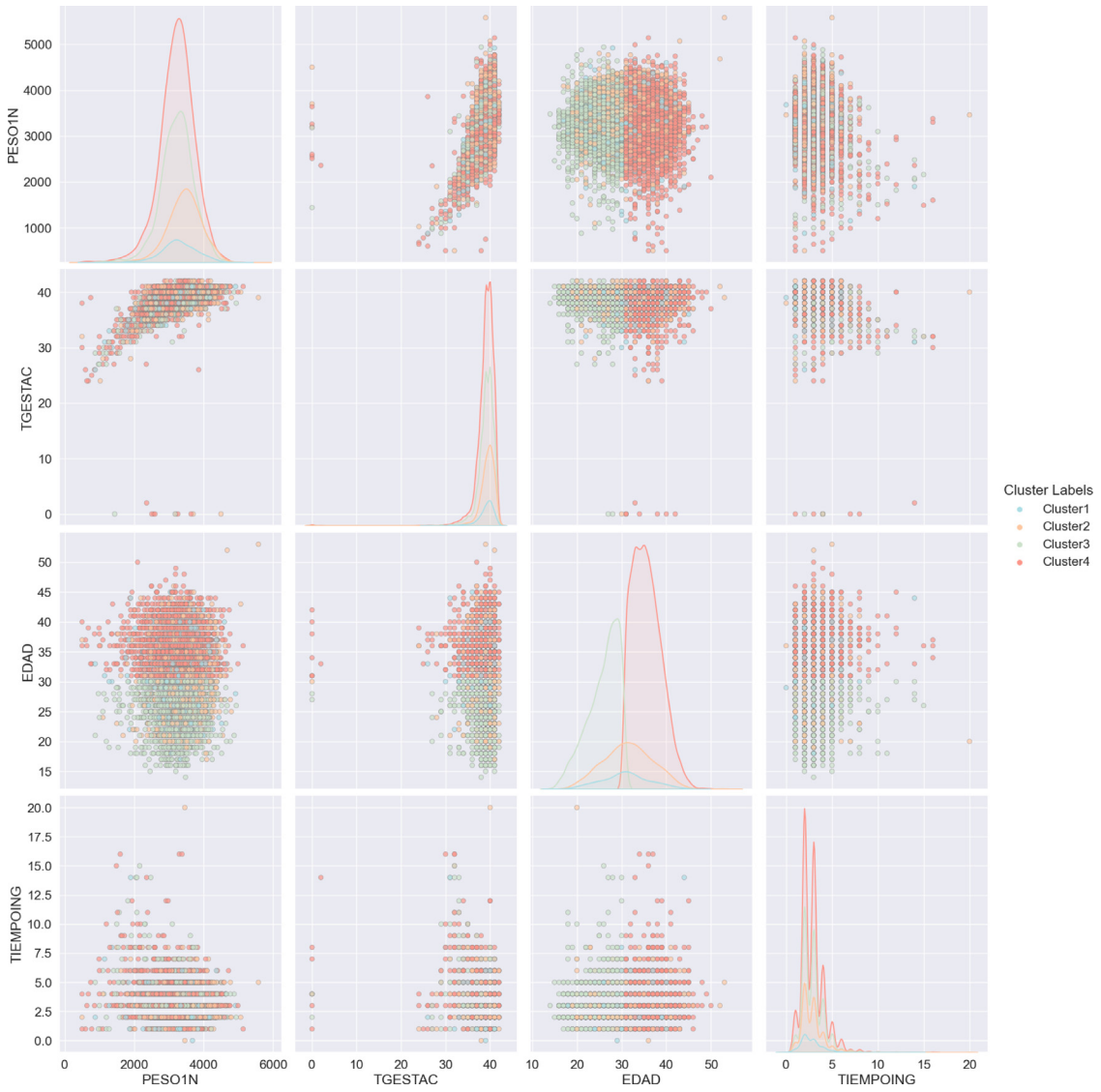



**Fig. 3.** Pairplot of the numerical variables. Two variables scatterplot and density plot in the main diagonal.

**Table 8**
Results interpretation.

| Cluster labels | Total individuals | Weight of the baby | Gestation time | Mother's age | Hospital admission time | Mother's country of birth |
|---|---|---|---|---|---|---|
| Cluster 1 | 391 | 3274.50 | 39.00 | 30.90 | 2.73 | UK |
| Cluster 2 | 1234 | 3444.69 | 39.29 | 31.45 | 2.77 | Morocco |
| Cluster 3 | 2463 | 3234.20 | 39.03 | 25.80 | 2.87 | Spain |
| Cluster 4 | 4016 | 3228.36 | 38.96 | 35.57 | 2.86 | Spain |

In addition to this example, we can see other results of algorithms implemented in the system in the state of the art. Among them, we can highlight works with fuzzy association rules in Big data for the extraction of hidden knowledge related to comorbidity in diagnoses [39].

## 6. Conclusions and future research

### 6.1. Future challenges

Through the platform proposed in this paper and the use case developed, it has been possible to see how the use of the Big Data platform can help in different processes involving various aspects of data management. Specifically, in our use case, we have seen how it improves knowledge extraction from data silos. Among the main advantages of our system are its application for managing massive amounts of data, that it is a user-friendly system and its ability to import heterogeneous data from different systems. Furthermore, to complete this capacity, integration with third-party IaaS will be implemented in order to be able to work in a hybrid way and improve the capacity of the platform.

However, as demonstrated in the use case, a large amount of data is not used due to security and anonymity constraints. In this sense, several approaches can be used to tackle this problem, such as federated learning, allowing the extraction of knowledge from data in a decentralized way. We can also include an additional layer that takes care of this process, thus preserving the modularity of the platform and facilitating its development and

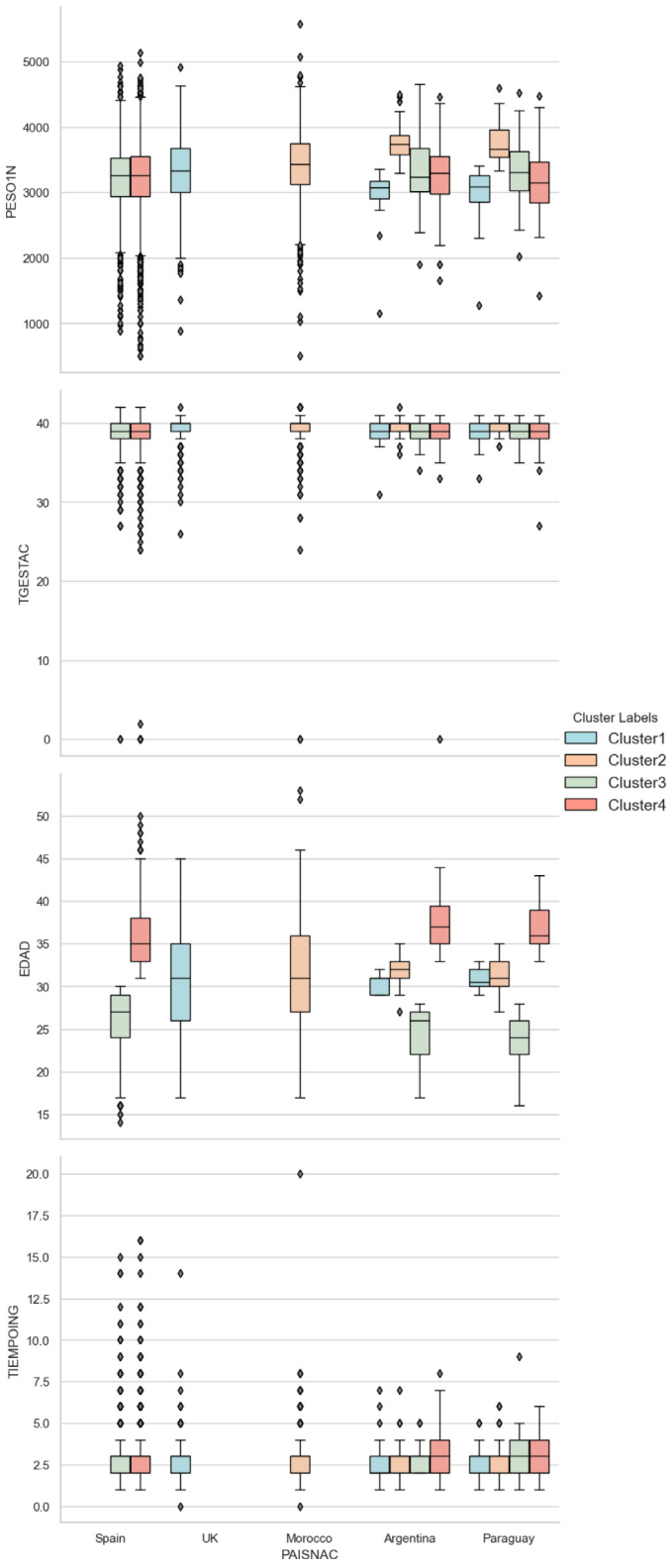



**Fig. 4.** Boxplot of the numerical and categorical variables. Numerical variables as rows and categorical as columns.

maintenance. However, we must remark that the data have been anonymized beforehand by the Andalusian Health System (SAS by its acronym in Spanish) for the use case presented. These approaches help maintain data anonymity and exploit hidden knowledge.

Regarding the integration of heterogeneous data, new methods have been implemented for the enrichment, and 'structuring' of unstructured data [29]. In addition, these methods pursue the objective of detrimental response time, which makes them more suitable for Big Data environments such as the one proposed in this article. Working in this direction will provide the AIMDP with greater flexibility, and further progress can be made in one of the platform's strong points: the treatment of heterogeneous data sources.

Other challenges that arise are the integration of data through data flows that can be found in a multitude of applications such as health, energy, and social networks. For this, it is necessary to adapt the integration of this type of data, and its processing and to have algorithms that allow the analysis of trends [62].

Finally, we have seen how a friendly and intuitive system for the end user improves the system's understanding, use and productivity. However, there are still challenges in the field of the use of data science tools by non-expert users in this field. Therefore, as enhancements to the system, an explainable artificial intelligence (XAI) layer has to be implemented to improve the interpretability and explainability of many of the methods that are available in our tool.

Using this future module, end-users will better understand the results obtained. Moreover, it will also be possible to improve the user experience by improving the generation and parameters of the algorithms by adding an intelligent AutoML process to select the best parameters and algorithms for users to obtain the desired results.

### 6.2. Conclusions

The proposed Modern Data Platform, AIMDP, has been effectively introduced in a real use case of a healthcare data silo, explaining its main capabilities. It has been demonstrated that it is possible to have a system capable of allowing a non-expert user to extract hidden knowledge using innovative technologies such as Big Data. Furthermore, this system has already included routines for data import [63], processing, data enrichment [39] and AI techniques to extract knowledge from large datasets [64–67]. The system offers a suitable framework for efficient and fault-tolerant data management due to the robustness of the systems used, such as micro-services, Spark etc.

Data is the heart of a system, although it is only stored in many cases, and the full knowledge it holds is not extracted. This is why platforms such as the one presented in this paper are necessary for the improvement of users in the different fields of energy, health, and economy to be able to analyse their data without a complex process in addition to being able to use innovative technologies and large computing systems in a transparent way.

This research opens the door to new implementations, improvements and applications of AIMDP to different fields, taking advantage of big data platforms and using the available Data Mining and Machine Learning algorithms that the scientific community has been developing in recent years. The main objective is to bring end users closer to the use of these new tools that can effectively help to manage and also to process the large volumes of data generated by social networks, medical records, images, sensors and other information external to the system, taking advantage of distributed computing and artificial intelligence in a simple, transparent and guided process.

## CRediT authorship contribution statement

**Alberto S. Ortega-Calvo:** Supervision, Original draft preparation, Investigation, Software, Writing, Visualization. **Roberto Morcillo-Jimenez:** Original draft preparation, Investigation,

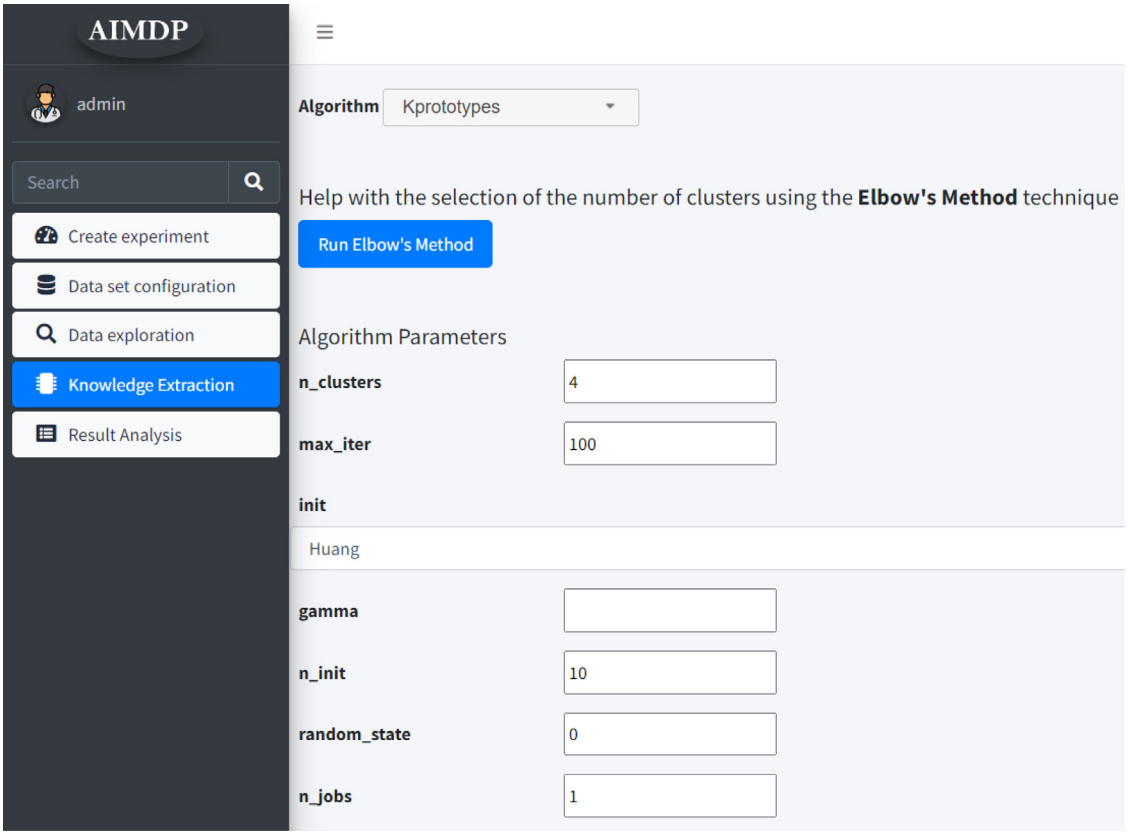



**Fig. 5.** Set up of the parameters of the K-Prototypes algorithm. Extracted from AIMDP data platform web-based application.

Software, Writing, Visualization. **Carlos Fernandez-Basso:** Conceptualization, Supervision,Writing – review & editing. **Karel Gutiérrez-Batista:** Conceptualization, Writing – review & editing. **Maria-Amparo Vila:** Supervision, Resources, Investigation, Writing – review & editing. **Maria J. Martin-Bautista:** Supervision, Resources, Investigation, Writing – review & editing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

The authors do not have permission to share data.

## Acknowledgements

The research reported in this paper was partially supported by the BIGDATAMED project, which has received funding from the Andalusian Government, Spain (Junta de Andalucía) under grant agreement No P18-RT-1765. In addition, this research has been partially supported by the Ministry of Universities through the EU-funded Margarita Salas programme NextGenerationEU.

## References

[1] M. Tanifuji, A. Matsuda, H. Yoshikawa, Materials data platform - a FAIR system for data-driven materials science, in: 2019 8th International Congress on Advanced Applied Informatics, IIAI-AAI, 2019, pp. 1021–1022, http://dx.doi.org/10.1109/IIAI-AAI.2019.00206.

[2] I. Vieira, A. Alvaro, A centralized platform of open government data as support to applications in the smart cities context, ACM SIGSOFT Softw. Eng. Notes 42 (4) (2018) 1–13.

[3] Y. Liu, J. Peng, Z. Yu, Big data platform architecture under the background of financial technology: In the insurance industry as an example, in: Proceedings of the 2018 International Conference on Big Data Engineering and Technology, 2018, pp. 31–35.

[4] B. Cheng, S. Longo, F. Cirillo, M. Bauer, E. Kovacs, Building a big data platform for smart cities: Experience and lessons from santander, in: 2015 IEEE International Congress on Big Data, IEEE, 2015, pp. 592–599.

[5] M.D. Ruiz, J. Gomez-Romero, C. Fernandez-Basso, M.J. Martin-Bautista, Big data architecture for building energy managament systems, IEEE Trans. Ind. Inform. (2021).

[6] X. Fei, K. Li, W. Yang, K. Li, Analysis of energy efficiency of a parallel AES algorithm for CPU-GPU heterogeneous platforms, Parallel Comput. 94 (2020) 102621.

[7] S. Denaxas, A. Gonzalez-Izquierdo, K. Direk, N.K. Fitzpatrick, G. Fatemifar, A. Banerjee, R.J. Dobson, L.J. Howe, V. Kuan, R.T. Lumbers, et al., UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER, J. Am. Med. Inform. Assoc. 26 (12) (2019) 1545–1559.

[8] University College London, CALIBER data platform official website, 2022, https://www.ucl.ac.uk/health-informatics/research/caliber/accessing-caliber-resources. (Accessed 14 June 2022).

[9] Y. Li, C. Wu, L. Guo, C.-H. Lee, Y. Guo, Wiki-health: A big data platform for health sensor data management, in: Cloud Computing Applications for Quality Health Care Delivery, IGI Global, 2014, pp. 59–77.

[10] T. Kariotis, M.P. Ball, B.G. Tzovaras, S. Dennis, T. Sahama, C. Johnston, H. Almond, A. Borda, Emerging health data platforms: From individual control to collective data governance, Data & Policy 2 (2020).

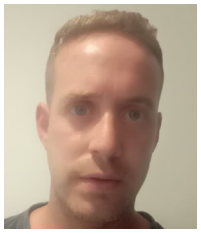
[11] PatientsLikeMe, PatientsLikeMe official website, 2022, https://www.patientslikeme.com/. (Accessed 17 June 2022).

[12] C. Fernandez-Basso, A.J. Francisco-Agra, M.J. Martín-Bautista, M.D. Ruiz, Finding tendencies in streaming data using Big Data frequent itemset mining, Knowl.-Based Syst. 163 (2019) 666–674, http://dx.doi.org/10.1016/j.knosys.2018.09.026.

[13] C. Fernandez-Basso, M.D. Ruiz, M.J. Martín-Bautista, A fuzzy mining approach for energy efficiency in a Big Data framework, IEEE Trans. Fuzzy Syst. 28 (11) (2020) 2747–2758, http://dx.doi.org/10.1109/TFUZZ.2020.2992180.

[14] C. Fernandez-Basso, M.D. Ruiz, M.J.M. Bautista, Spark solutions for discovering fuzzy association rules in Big Data, Internat. J. Approx. Reason. 137 (2021) 94–112, http://dx.doi.org/10.1016/j.ijar.2021.07.004.
[15] K. Gutiérrez-Batista, J.R. Campaña, M.A.V. Miranda, M.J. Martín-Bautista, An ontology-based framework for automatic topic detection in multilingual environments, Int. J. Intell. Syst. 33 (7) (2018) 1459–1475, http://dx.doi.org/10.1002/int.21986.
[16] K. Gutiérrez-Batista, J.R. Campaña, M.A.V. Miranda, M.J. Martín-Bautista, Building a contextual dimension for OLAP using textual data from social networks, Expert Syst. Appl. 93 (2018) 118–133, http://dx.doi.org/10.1016/j.eswa.2017.10.012.
[17] K. Gutiérrez-Batista, M.A. Vila, M.J. Martín-Bautista, Building a fuzzy sentiment dimension for multidimensional analysis in social networks, Appl. Soft Comput. 108 (2021) 107390, http://dx.doi.org/10.1016/j.asoc.2021.107390.
[18] A. Helmond, The platformization of the web: Making web data platform ready, Soc. Media + Soc. 1 (2) (2015) 2056305115603080.
[19] D. Zburivsky, L. Partner, Designing Cloud Data Platforms, Simon and Schuster, 2021.
[20] X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde, S. Owen, et al., Mllib: Machine learning in apache spark, J. Mach. Learn. Res. 17 (1) (2016) 1235–1241.
[21] A. Spark, Apache spark, 17, (1) 2018, p. 2018, Retrieved January.
[22] L. Hirth, J. Mühlenpfordt, M. Bulkeley, The ENTSO-E transparency platform–A review of Europe's most ambitious electricity data platform, Appl. Energy 225 (2018) 1054–1067.
[23] ENTSO-E, ENTSO-E transparency platform official website, 2022, https://www.entsoe.eu/data/transparency-platform/. (Accessed 14 June 2022).
[24] C. Scheidt-Nave, P. Kamtsiuris, A. Gößwald, H. Hölling, M. Lange, M.A. Busch, S. Dahm, R. Dölle, U. Ellert, J. Fuchs, et al., German health interview and examination survey for adults (DEGS)-design, objectives and implementation of the first data collection wave, BMC Pub. Health 12 (1) (2012) 1–16.
[25] X. Wang, Y. Zhang, V.C. Leung, N. Guizani, T. Jiang, D2D big data: Content deliveries over wireless device-to-device sharing in large-scale mobile networks, IEEE Wirel. Commun. 25 (1) (2018) 32–38.
[26] X. Hu, M. Yuan, J. Yao, Y. Deng, L. Chen, Q. Yang, H. Guan, J. Zeng, Differential privacy in telco big data platform, Proc. VLDB Endow. 8 (12) (2015) 1692–1703.
[27] N. Luo, M. Pritoni, T. Hong, An overview of data tools for representing and managing building information and performance data, Renew. Sustain. Energy Rev. 147 (2021) 111224.
[28] MongoDB, What is a data platform? 2021, https://www.mongodb.com/what-is-a-data-platform. (Accessed 31 May 2022).
[29] F. Cauteruccio, P.L. Giudice, L. Musarella, G. Terracina, D. Ursino, L. Virgili, A lightweight approach to extract interschema properties from structured, semi-structured and unstructured sources in a big data scenario, Int. J. Inf. Technol. Decis. Mak. 19 (03) (2020) 849–889.
[30] J. Chen, N. Yang, M. Zhou, Z. Zhang, X. Yang, A configurable deep learning framework for medical image analysis, Neural Comput. Appl. 34 (10) (2022) 7375–7392.
[31] P. Mell, T. Grance, et al., The NIST Definition of Cloud Computing, Computer Security Division, Information Technology Laboratory, National …, 2011.
[32] M.D. Assunção, R.N. Calheiros, S. Bianchi, M.A. Netto, R. Buyya, Big data computing and clouds: Trends and future directions, J. Parallel Distrib. Comput. 79 (2015) 3–15.
[33] P.A. Forero, A. Cano, G.B. Giannakis, Consensus-based distributed support vector machines, J. Mach. Learn. Res. 11 (5) (2010).
[34] J. Chen, K. Li, Q. Deng, K. Li, S.Y. Philip, Distributed deep learning model for intelligent video surveillance systems with edge computing, IEEE Trans. Ind. Inform. (2019).
[35] CPRD, Clinical practice research datalink official website, 2022, https://cprd.com/. (Accessed 14 June 2022).
[36] G.A. Williams, S.M.U. Díez, J. Figueras, S. Lessof, et al., Translating evidence into policy during the COVID-19 pandemic: bridging science and policy (and politics), Eurohealth 26 (2) (2020) 29–33.
[37] V. Palanisamy, R. Thirunavukarasu, Implications of big data analytics in developing healthcare frameworks–A review, J. King Saud Univ. Comput. Inf. Sci. 31 (4) (2019) 415–425.
[38] C.S. Kruse, A. Stein, H. Thomas, H. Kaur, The use of electronic health records to support population health: a systematic review of the literature, J. Med. Syst. 42 (11) (2018) 1–16.
[39] C. Fernandez-Basso, K. Gutiérrez-Batista, R. Morcillo-Jiménez, M.-A. Vila, M.J. Martin-Bautista, A fuzzy-based medical system for pattern mining in a distributed environment: Application to diagnostic and co-morbidity, Appl. Soft Comput. 122 (2022) 108870.
[40] J. Waring, C. Lindvall, R. Umeton, Automated machine learning: Review of the state-of-the-art and opportunities for healthcare, Artif. Intell. Med. 104 (2020) 101822.
[41] E. LeDell, S. Poirier, H2O automl: Scalable automatic machine learning, in: Proceedings of the AutoML Workshop At ICML, Vol. 2020, 2020.
[42] B. Raef, R. Ferdousi, A review of machine learning approaches in assisted reproductive technologies, Acta Inform. Medica 27 (3) (2019) 205.
[43] W. McKinney, pandas: powerful Python data analysis toolkit, 2008, URL https://pandas.pydata.org/.
[44] Oracle, Oracle corporation, 1977, URL https://www.oracle.com/.
[45] GitHub, GitHub, inc., 2008, URL https://github.com/.
[46] M. Grinberg, Flask API-RESTful, 2013, URL https://flask-api-restful.readthedocs.io/en/latest/.
[47] MongoDB, Inc., MongoDB, 2007, URL https://www.mongodb.com/.
[48] Docker, Docker official website, 2022, https://www.docker.com/. (Accessed 27 June 2022).
[49] Amazon.com, Inc., Amazon web services, 2006, URL https://aws.amazon.com/.
[50] Google, Google LLC, 1998, URL https://www.google.com/.
[51] Microsoft Corporation, Microsoft azure, 2010, URL https://azure.microsoft.com/.
[52] W.H. Inmon, OLAP cubes, Commun. ACM 39 (9) (1996) 90–98, http://dx.doi.org/10.1145/237257.237273.
[53] C. Fernandez-Basso, M.D. Ruiz, M.J. Martin-Bautista, Extraction of association rules using big data technologies, Int. J. Des. Nat. Ecodynamics 11 (3) (2016) 178–185.
[54] Z. Zhang, J. Jiang, W. Wu, C. Zhang, L. Yu, B. Cui, Mllib*: Fast training of glms using spark mllib, in: 2019 IEEE 35th International Conference on Data Engineering, ICDE, IEEE Computer Society, 2019, pp. 1778–1789.
[55] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, et al., Xgboost: extreme gradient boosting, 1, (4) 2015, pp. 1–4, R package version 0.4-2.
[56] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[57] M.A.B. HajKacem, C.E.B. N'Cir, N. Essoussi, KP-S: a spark-based design of the K-prototypes clustering for big data, in: 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications, AICCSA, IEEE, 2017, pp. 557–563.
[58] J. Kim, H. Ha, B.-G. Chun, S. Yoon, S.K. Cha, Collaborative analytics for data silos, in: 2016 IEEE 32nd International Conference on Data Engineering, ICDE, IEEE, 2016, pp. 743–754.
[59] C. Zhang, S. Li, J. Xia, W. Wang, F. Yan, Y. Liu, {BatchCrypt}: Efficient homomorphic encryption for {Cross-Silo} federated learning, in: 2020 USENIX Annual Technical Conference, USENIX ATC 20, 2020, pp. 493–506.
[60] SAS, Minimum basic data set (MBDS) of Andalusia, 2022, https://www.sspa.juntadeandalucia.es/servicioandaluzdesalud/profesionales/sistemas-de-informacion/cmbd-andalucia. (Accessed 20 June 2022).
[61] WHO, ICD-10 - International statistical classification of diseases and related health problems 10th revision, 2022, https://icd.who.int/browse10/2019/en#/. (Accessed 27 June 2022).
[62] C. Fernandez-Basso, A.J. Francisco-Agra, M.J. Martin-Bautista, M.D. Ruiz, Finding tendencies in streaming data using big data frequent itemset mining, Knowl.-Based Syst. 163 (2019) 666–674.
[63] C. Fernandez-Basso, M.D. Ruiz, M.J. Martin-Bautista, A fuzzy mining approach for energy efficiency in a Big Data framework, IEEE Trans. Fuzzy Syst. 28 (11) (2020) 2747–2758.
[64] M.D. Ruiz, D. Sánchez, M. Delgado, M.J. Martin-Bautista, Discovering fuzzy exception and anomalous rules, IEEE Trans. Fuzzy Syst. 24 (4) (2015) 930–944.
[65] K. Gutiérrez-Batista, J.R. Campaña, M.-A. Vila, M.J. Martin-Bautista, An ontology-based framework for automatic topic detection in multilingual environments, Int. J. Intell. Syst. 33 (7) (2018) 1459–1475.
[66] I. Diaz-Valenzuela, V. Loia, M.J. Martin-Bautista, S. Senatore, M.A. Vila, Automatic constraints generation for semisupervised clustering: experiences with documents classification, Soft Comput. 20 (6) (2016) 2329–2339.
[67] C. Fernandez-Basso, M.D. Ruiz, M.J. Martin-Bautista, Spark solutions for discovering fuzzy association rules in Big Data, Internat. J. Approx. Reason. 137 (2021) 94–112.

**Alberto S. Ortega-Calvo** received his Computer Engineering degree and Master's degree in Data Science in 2020 and 2021, respectively. He completed both at the University of Granada, for which he is currently actively researching together with the IdBIS (Intelligent Databases and Information Systems) research group on projects such as P18-RT-1765, whose main topics are Big Data, medical data analysis and machine learning. He is currently posing a thesis on topics related to medical data processing and federated learning.



**Roberto Morcillo Jiménez** received a degree in Computer Engineering and M. in Computer Engineering from the University of Granada, Spain. He worked as an Assistant Professor in the Department of Computer Science and Artificial Intelligence between 2016 and 2019. He is doing a Ph.D. in Computer Science and Artificial Intelligence at the University of Granada from 2019. He is currently working as an Assistant Professor in the Department of Computer Languages and Systems. He is Associated Research of Intelligent Data Bases and Information Systems (IDBIS) research group at the University of Granada. His research interests comprise Deep Learning, Reinforcement Learning, Building Energy Efficiency and Control, and Information Systems.

**Carlos Fernandez-Basso** received the degree in computer science, the M.Sc. degree in data science, and the Ph.D. degree in computer science from the University of Granada, Granada, Spain, in 2014, 2015, and 2020, respectively.,He is currently a Postdoctoral Fellow with Causal Cognition Lab, University College London, London, U.K. He was a Lead Developer in the EU FP7 Project Energy IN TIME in the topics of building simulation and control, data analytics, and machine learning, and in the COPKIT Project in the topics of cybercrime, Big Data, and machine learning. From 2016 to 2018, he collaborated with the Data Science Institute, Imperial College London, London, U.K., where he has carried out research stays.

**Karel Gutiérrez-Batista** received a degree in computer science and an M.Sc. degree in data science from the University of Camagüey, Cuba. He worked as an Assistant Professor between 2009 and 2014 at the Department of Computer Science at the University of Camagüey. He

received his Ph.D. in computer science in 2018 from the University of Granada, Spain. He is currently working as a Postdoctoral Fellow in the Department of Computer Science and Artificial Intelligence at the University of Granada, Spain. He is an Associated Research of Intelligent Data Bases and Information Systems (IDBIS) research group at the University of Granada. His research interests comprise Multidimensional Data Analysis, Deep Learning, Data Mining, and Natural Language Processing.

**Maria-Amparo Vila Miranda** is Professor of Computer Science and Artificial Intelligence, University of Granada since 1992, she was previously associate professor and analystprogrammer of the Computer Centre at the same university. She has developed her research and teaching primarily in the area of databases and intelligent information systems, and its main lines of research are : treatment of imprecision in information systems using fuzzy logic, knowledge discovery in databases using techniques "Soft Computing"-based, and ubiquitous computing and knowledge mobilization. She has been the advisor of 27 doctoral theses, she has been or is responsible for more than 10 research projects and she has published numerous research papers, which highlights more than 100 papers in SCI journals. She has been responsible for the research group of the Andalusian ICT-113 Approximate Reasoning and Artificial Intelligence) from 1994 to 1997 and head of the group TIC174 (Databases and Intelligent Information Systems) since its inception in 2000 until today . From the point of view of university management she has been Vice-Director of Organization at the School of Computer Science for 3 years (1994-1997) and Director of the Department of Computer Science and Artificial Intelligence, University of Granada from 1997 to 2004. She has also developed related tasks and quality assessment processes in this regard has been: a member of the speech area of Information Technology and Communications in the Andalusian Research Plan III, II and evaluator within the Quality Plan Universities (University Council ersities MEC) has been evaluating also for ANECA within the institutional evaluation program, having served as president of various committees of external and Commissioner Assessment Andalusian Autonomous Accessories (CAECA) being responsible and chairwoman of the subcommittee of the area of Technical Education's. She has been also chairman of the evaluation committee of undergraduate Architecture and Engineering within the program VERIFICA of ANECA. This committee is responsible for assessments and guidelines teachers of all grades of Computer Engineering and Telecommunications have been evaluated in Spain.

**Dr. Maria J. Martin-Bautista** is a Full Professor at the Department of Computer Science and Artificial Intelligence at the University of Granada, Spain, since 1997. She is a member of the IDBIS (Intelligent Data Bases and Information Systems) research group. Her current research interests include Data Science and Big Data Analytics in Data, Text, Web and Social Networks, Intelligent Information Systems, Knowledge Representation and Uncertainty. She has supervised several Ph.D. Thesis and published more than 100 papers in high impact international journals and conferences. She has participated in more than 20 R+D projects and has supervised several research technology transfers with companies. She has served as a program committee member for several international conferences.