Co-funded by the European Union

FAAI:
The Future is in Applied Artificial Intelligence
Erasmus+ project 2022-1-PL01-KA220-HED-000088359

01.09.2022 – 31.08.2024

Research 7: Collecting IT specifications of good practices in AI:
the state-of-the-art analysis for WP2

Co-funded by the European Union

The production of this document has been possible thanks to the support of the ERASMUS+ project: The Future is in Applied Artificial Intelligence (2022-1-PL01-KA220-HED-000088359)

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the National Agency (NA). Neither the European Union nor NA can be held responsible for them.

Date
15.03.2023

Places of the development of the result
University of Bielsko-Biala, Bielsko-Biala, Poland
University of Library Studies and Information Technologies, Sofia, Bulgaria
University of Nis, Serbia
University of Ss. Cyril and Methodius in Trnava, Slovakia
University of Montenegro, Montenegro

Summary: The work presents the study of specifications of good practices in applied artificial intelligence (AAI). The analysis of 25 questionnaires from five partner institutions revealed key insights into the current state of artificial intelligence (AI) and machine learning (ML) projects. Training conducted in Serbia and Bulgaria, was signaling a need for expanded opportunities in EU countries. As a result of the study, we obtained that Deep ML prevails, particularly in Convolutional Neural Networks, while Gated Recurrent Unit is less common. Data volumes between 1 GB and 1 TB are typical, reflecting practical constraints. AI applications span diverse fields, with TensorFlow leading in libraries. Permissive licenses are most prevalent, databases are primary data sources, and texts/pictures dominate data characteristics. NoSQL databases are favored for storage. Security features and data processing tools vary. Dedicated servers and clusters are widely used, recommender systems are prominent, Python is the preferred language, and Apache Hadoop dominates ecosystems. Free datasets foster accessibility. Overall, the findings emphasize the dynamic nature of AI/ML projects, providing a foundation for future research in the rapidly advancing field.

Keywords: applied artificial intelligence, machine learning, good practice, convolutional neural networks

1. Introduction

Artificial intelligence (AI) and machine learning (ML) have become integral components of cutting-edge research, driving innovation and transformative solutions across diverse industries. In a comprehensive analysis, we delve into the landscape of AI and ML projects undertaken by various research institutions. This exploration, based on data collected from 25 questionnaires across five partner institutions, unveils critical insights that shape the current state of AI and ML initiatives.

The findings reveal a rich tapestry of trends, challenges, and opportunities within the realm of AI. From training dynamics in Serbia and Bulgaria to the prevalent use of Deep Machine Learning models, the survey outlines a nuanced picture of the AI landscape. As we navigate through the key outcomes, it becomes evident that the applications of AI span across multifaceted domains, including healthcare, finance, smart cities, and Industry 4.0.

TensorFlow emerges as a frontrunner among AI libraries, underscoring the importance of tool selection tailored to project-specific needs. The prevalence of Convolutional Neural Networks (CNNs) in diverse applications, alongside the scarcity of certain ML models like SciMl, reflects the evolving preferences and challenges in the field.

Data considerations, from processing volumes between 1 GB and 1 TB to the dominance of free datasets, provide insights into the practicalities and accessibility shaping AI endeavors. The prominence of permissive licenses, predominant use of

NoSQL databases, and the nuanced nature of security feature implementation further characterize the intricate landscape of AI projects.

As we traverse through the details of data characteristics, processing methodologies, and the varied outcomes achieved, it becomes evident that AI is not a one-size-fits-all domain. The dynamic nature of AI projects demands tailored solutions and continuous adaptation to emerging technologies.

The objective of the study is a foundational exploration into the dynamic and evolving field of AI. The identified trends and patterns act as guideposts for future research and development, emphasizing the need for adaptability and innovation in the rapidly advancing realm of artificial intelligence.

2. Collection and analysis of data

The data was acquired by five partner institutions scientists. In the current research 8 questionnaires were obtained from UBB, 6 questionnaires from ULSIT, 5 from UNi, 5 from UoM and 1 from USCM.
In total 25 questionnaires were collected by 11 researchers.

3. Results

3.1. The country in which the training takes place

The first question of the survey, after the data for an organization name and a researcher, asks where the training takes place. The results are presented below.

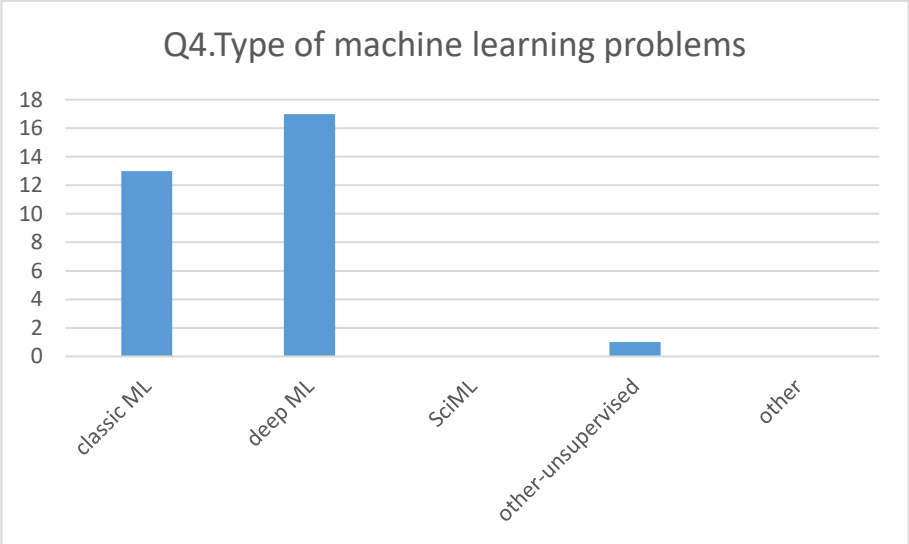
Data description:
According to the research data, the training takes place mostly in Serbia, followed by Bulgaria. One respondent answered that they only used Coursera or O'Reilly for trainings. Between one and three trainings take place in the other countries in the survey.

Discussion:
The survey shows that there is a need for more trainings for IT specialists so that more good practices in the field of AI can be observed. **Main conclusions:**

- There is a need for more trainings in the EU countries.
- The most trainings take place in Serbia and Bulgaria.

3.2. Type of machine learning problems

The next question reveals the type of machine learning models which have been used.



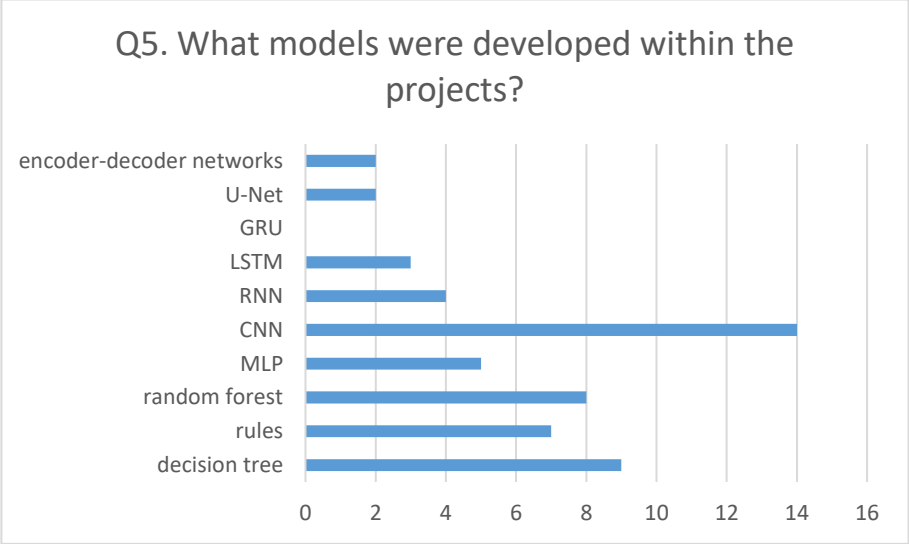
Data description:
The data shows that the most used machine learning model is deep ML, followed by classic ML. Over the past years deep learning has become a major tool for a variety of AI problems. It has demonstrated better performance on different tasks, including natural language processing, vision, virtual assistants, chatbots, healthcare, etc.

Discussion:

- The most used model, based on the survey, is deep ML.
- SciMl or other machine learning models are not common among the projects reviewed.

3.3. What models were developed (studied) within the projects?

The following question asks what models were studied in particular. The results are shown in the chart below.



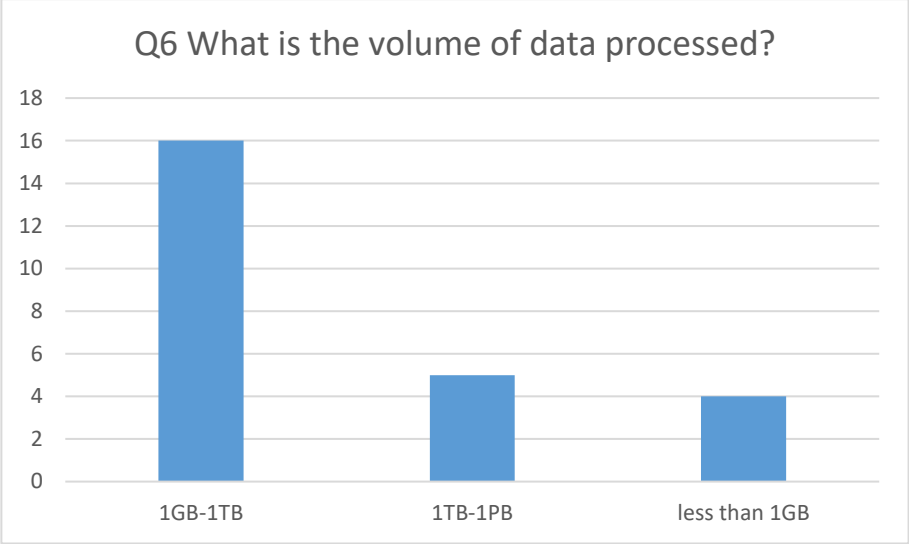
Data description:
According to the results the models used the most in the projects were the convolutional neural networks. It is understandable because those models have diverse applications – image classification, object detection, facial recognition, medical image analysis, NLP, autonomous vehicles or technology. CNNs are particularly appropriate for image recognition tasks because they can automatically learn to detect complex features such as edges, corners, and textures. Decision tree, together with random forest, are other models which were developed for different AI problems. Many applications of decision trees and random forests, are widely used in industries such as finance, healthcare, manufacturing and environmental monitoring. Decision trees and random forests are popular machine learning algorithms because they are easy to use, understand and interpret, can handle both numerical and categorical data and can process large and complex datasets.

Discussion:

- Convolutional neural networks are the most common models in the projects.
- GRU was not developed within the projects.

3.4. What models were developed (studied) within the projects?

The next question points to the volume of data which was processed. The possible answers were less than 1GB, 1GB – 1TB, 1TB – 1PB and over 1 PB.



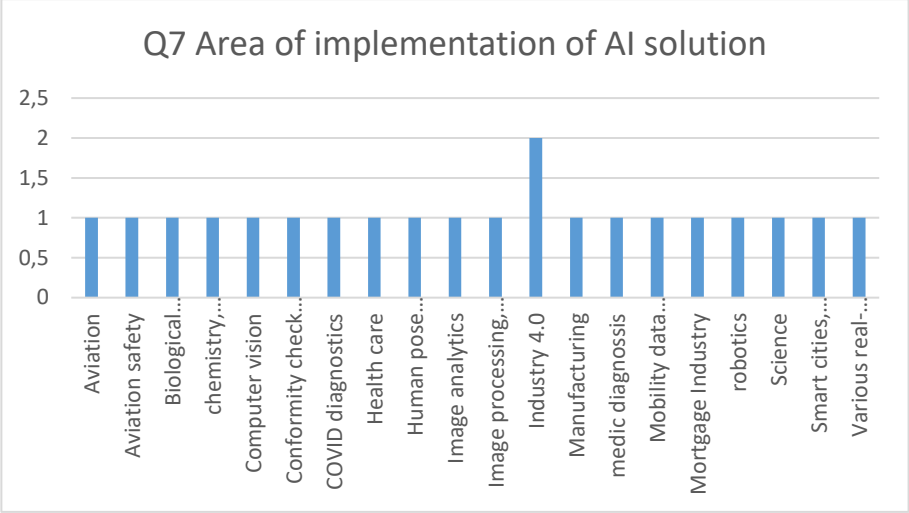
Data description:
As it can be seen in the chart, the volume of data processed was mostly between 1 GB and 1 TB. That volume represents relatively large amounts of data which are commonly encountered in many different types of projects, but are still manageable in terms of storage and processing capabilities of the most modern computers. The volume of data processed more than 1 TB is not so common, probably due to accessibility and cost. Managing and processing data at the petabyte scale requires specialized infrastructure and resources that are not readily available to most organizations. Storing and processing large amounts of data can be expensive, both in terms of hardware and operational costs. On the other hand volume, less than 1 GB, is not common either because it is considered to be small enough to be processed efficiently on a single machine, without the need for specialized hardware or distributed computing systems.

Discussion:

- The volume of data processed between 1 GB and 1 TB is used the most.
- Volume more than 1 TB or less than 1 GB is not so applicable.

3.5. What is the area of implementation of AI solution?

The next question is directed to areas of implementation of AI solutions. The answers include respondents' own input, based on each case. The results are shown graphically below.



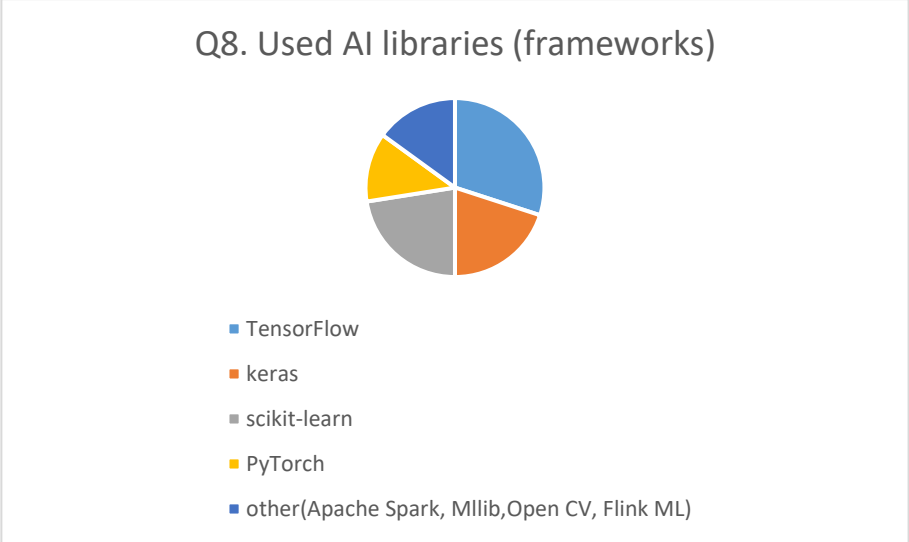
Data description:
It should be noted that the researchers looked for projects in different fields, which are as follows: agriculture, AI in medicine, surgery, air traffic management, aviation, aviation safety, biological sequence analysis, chemistry, robotics, health, computer vision, conformity check in aerospace industry, COVID diagnostics, health care, human pose estimation, image analytics, image processing, price prediction, industry 4.0, manufacturing, medical diagnosis, mobility data science and analytics, mortgage industry, science, smart cities, traffic monitoring, various real-world cases. Each respondent investigated one or two of these areas of implementation of AI solutions.

Discussion:

- Each respondent was assigned to explore a particular area so that as many areas as possible could be reviewed.

3.6. Used AI libraries (frameworks).

Question 8 suggests different AI libraries which are usually used in the domain of artificial intelligence. They are shown in the pie chart below.



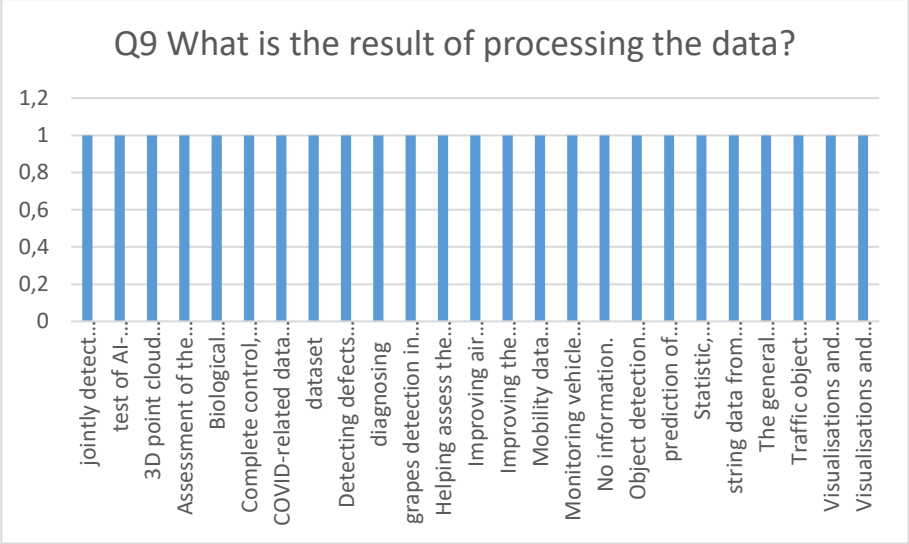
Data description:
The results indicate that the most used AI library is TensorFlow. TensorFlow is one of the most popular and widely used open-source libraries for building and deploying machine learning and deep learning models. The popularity of that library can also be attributed to its large and active community, which provides support and contributes to its development. Other popular AI libraries include Keras and Scikit-learn. The other libraries used in AI problems involve PyTorch, Apache Spark, MLlib, Open CV and Flink ML.

Discussion:

- The choice of which library to use depends on various factors, including the specific needs of the project, personal preferences, and the level of expertise of the developer.

3.7. What is the result of processing the data?

This question includes open answers of the respondents, depending on each case. Therefore, there were twenty-five different results for each project.



Data description:

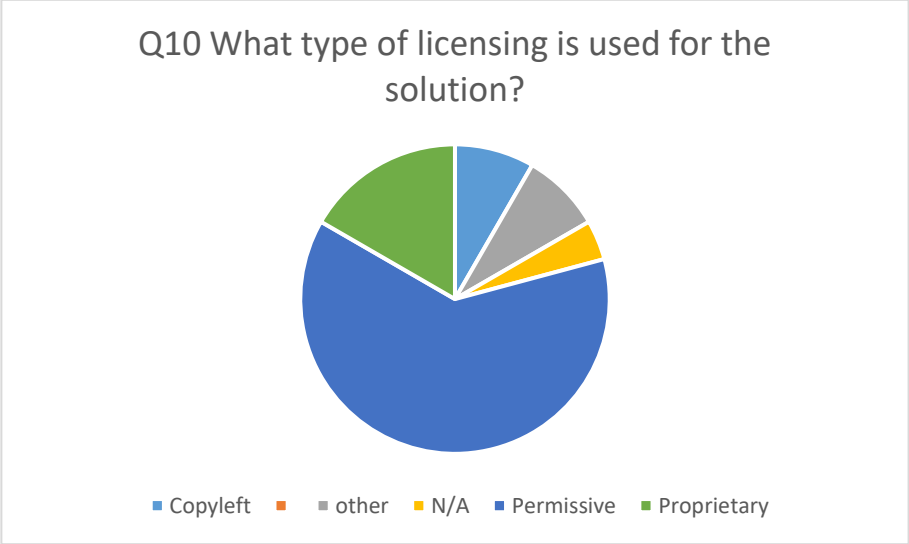
In the first project the result was joint detection of human body, facial and foot key points. In the second project the result was a test of AI-TWILIGHT methods in automotive, horticulture and street lighting domains. The third result was 3D point cloud classification and segmentation. The fourth result was assessment of the likelihood and the severity of the Covid-19 infection. The fifth result was biological sequence classification. The sixth result was complete control, full articulation, and intelligent feedback. The seventh result was COVID-related data analytics. The eighth result was creation of a dataset. The ninth result was directed to detecting defects on products in manufacturing. The tenth result was directed to implementation in diagnosis. The eleventh result was grapes detection in vineyard. The twelfth result was aid in assessment of the status of machinery and overall equipment within the factory. The thirteenth result was improvement in air traffic management through machine learning collaboration on private data sets. The fourteenth result was improvement of the decision-making process in go-around scenarios, which is of great importance for the safety of both airlines and air navigation service providers in ATM. The fifteenth result was mobility data analytics. The sixteenth result was monitoring of a vehicle which flows in cities by counting cars from images acquired from smart cameras. The seventeenth result was object detection and recognition. The eighteenth result was prediction of activities of chemical compounds, automatic control of robots, analysis of human movements from wearable sensors. The nineteenth result was statistics, prescribing medication and appointment of treatment. The twentieth result was obtaining data from images and object from images. The twenty-first result was an attempt to develop a prototype of a global multi-hazard monitoring and early warning system. The twenty-second result was traffic object detection and recognition. The twenty-third result presented visualisations and information about future prices. The twenty-fourth result presented visualisations and ML models. There was one project in which no results were documented.

Discussion:

- Each respondent had to find results after processing the data, depending on each project. In this way various outcomes in AI domains were discovered.

3.8. What type of licensing is used for the solution?

The following question asks what type of licensing is used in the solution and includes four alternatives – permissive (BSD, MIT), copyleft (GPL, LGPL), proprietary (Bespoke, Commercial) and other. The chart below shows the distribution of these types of licensing.



Data description:

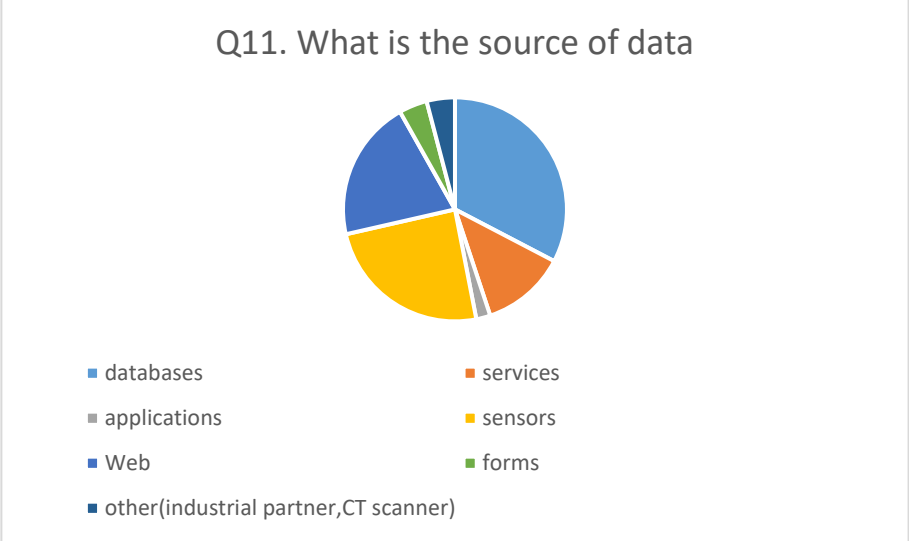
The results show that the most used type of licensing is the permissive one (in 15 projects), followed by the proprietary (in 4 projects), copyleft (in 2 projects). In one project a freely available for non-commercial use original License of OpenPose was used. Permissive licenses are popular because they offer a high degree of freedom to users and developers. These types of licenses typically allow users to modify and redistribute the software without requiring that any changes or improvements be released under the same license. This is in contrast to copyleft licenses, which require any derivative works to be licensed under the same terms as the original. Permissive licenses are often chosen by individuals and organizations that want to encourage collaboration and innovation, while at the same time allowing for maximum flexibility in how the software is used and distributed. Proprietary licensing takes the second position although it is found in four projects only. That type of licensing is common in the software industry but often limits the ways in which users can use and distribute the software. Some companies still choose to use proprietary licensing models as a way to protect their intellectual property and maintain control over their software. Copyleft licenses are not as common as permissive or proprietary licenses because they impose more restrictions on how software can be used and distributed

Discussion:

- Permissive licenses are popular because they offer a balance of freedom and flexibility that allows for collaboration and innovation while minimizing barriers to entry and adoption.
- Copyleft licenses are less common than permissive or proprietary licenses because they impose more restrictions, which may not be desirable for all users and developers.

3.9 What is the source of data?

Question 11 gives responses, regarding the source of data. The options vary from databases, services, applications, sensors, Web, forms or other (if the answer is other, the respondent inputs the source). The pie chart below shows the responses and their distribution.



Data description:

The results demonstrate that the basic source of data are the databases (in 16 reports), followed by sensors (in 12 reports) and Web (in 10 reports). The least common source are the forms (in 2 reports), other, such as industrial partner or CT scanner (in 2 reports) and applications (in 1 report). Databases are preferred sources because the data there are structured and can be retrieved efficiently. In addition the data are consistent and accurate and databases can easily integrate with other systems, making it easy to share information between different applications and platforms. On the other hand applications and forms are not a common source of information because the data are often scattered and their quality is inconsistent. Applications or forms may not be designed to handle large amounts of data and access may be restricted due to security or privacy concerns. In addition, Forms may use different data formats and protocols, which can make it challenging to integrate data across different systems and applications.

Discussion:

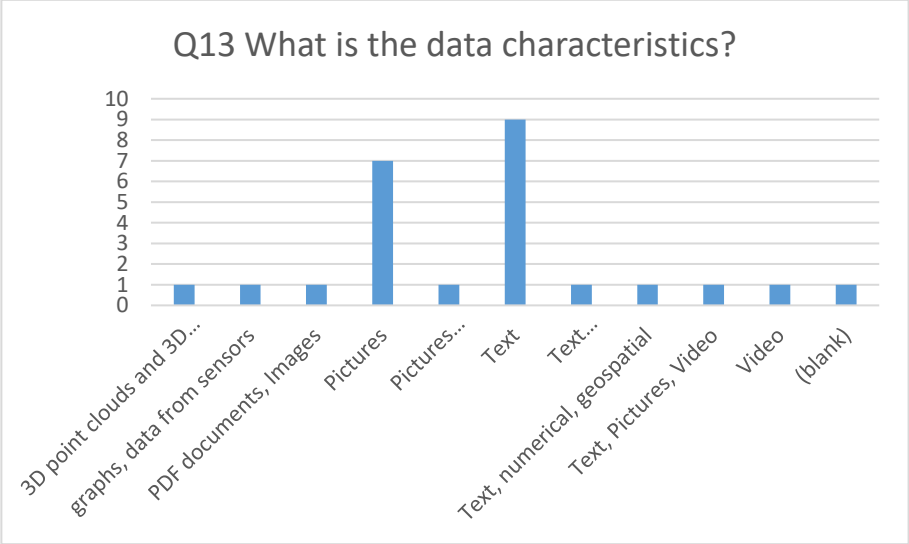
- Databases are a preferred source of data because they offer a reliable, efficient, and secure way to store and manage large amounts of data in consistent and structured manner.
- Applications and forms can be sources of data but they often require significant effort and resources to extract and use effectively.

3.10. Data representation.

This question gives information about the type of data. The following types are included in the projects - graphs, data from sensors, text, images, csv, avro, parquet, JSON API, MySQL server, FTP server, 3D point cloud, chest CT scans, medical research**.**

3.11. What is the data characteristics?

The following question gives information about the characteristics of the data. The possible answers are pictures, text, video, audio or other.



Data description:
As it can be seen in the chart, the main characteristics of the data are the texts and pictures or pictures with text or video. Pictures and texts are easy to understand and interpret. They can be generated in large volumes and can be used used in a variety of contexts, from marketing and advertising to scientific research and data analysis. Pictures and texts can be easily accessed and shared across different platforms.

Discussion:

- Pictures and texts are common data characteristics because they are easy to understand, generate in high volumes, versatile, and accessible. Advances in

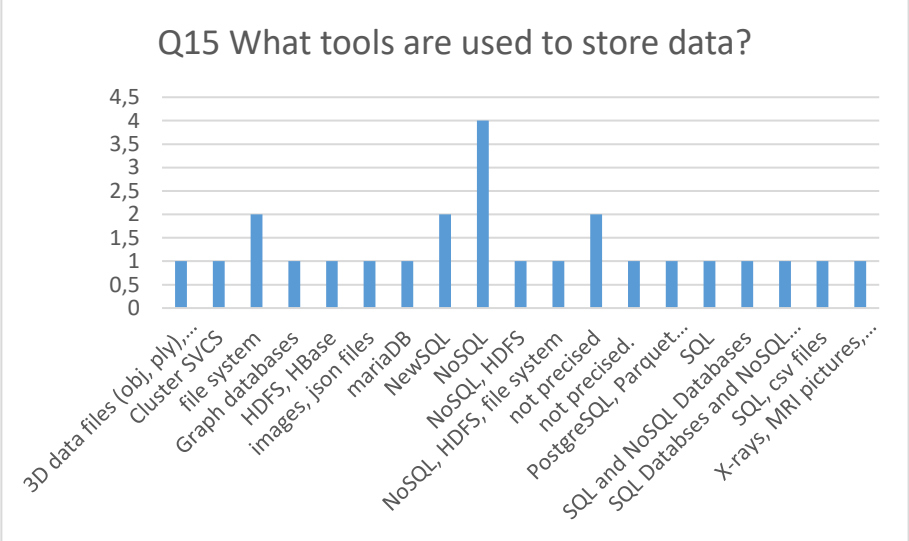
technology have also made it easier to analyze and extract insights from unstructured data, making them a valuable source of information for businesses and researchers alike.

3.12. Data processing and quality.

Question 14 gives information about the quality of the data and how the data have been processed. Different methods have been used, based on the case. In some projects expert decisions have been made based on medical expertise. In other projects files have been accessed and formatted via Python. In some projects data have been cleaned and visualised. In other projects classification and regression have been used. In some cases scaling, labelling, audio or video processing have been applied. These techniques were found to be the most common ones in data processing**.**

3.13. What tools are used to store data?

This question is open and gives information about the tools which are used to store data in the different projects. The chart below shows the responses.



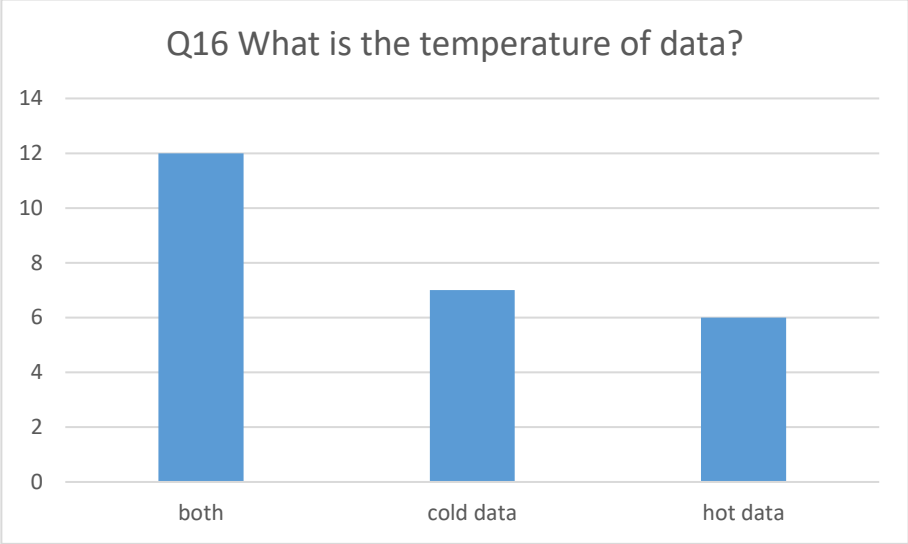
Data description:
The most common tools are NoSQL databases. Those databases are designed to handle unstructured or semi-structured data which makes them a better choice than traditional relational databases for applications that require flexible data modelling. This allows developers to store and retrieve data in a way that better suits the needs of their application. Many NoSQL databases are open-source, which makes them a cost-effective option for developers and organizations. The other tools are almost equally distributed, so it can be concluded that the tool which will be used depends on the case and IT problem.

Discussion:

- Overall, NoSQL databases offer several advantages over traditional relational databases and that is why they are becoming increasingly common as tools to store data. They offer more flexibility, scalability, availability, and performance, which makes them an ideal choice for modern applications that require these features.

3.14. What is the temperature of data?

The following questions reveals if cold, hot or both data are more common. The chart below shows the distribution of these types.



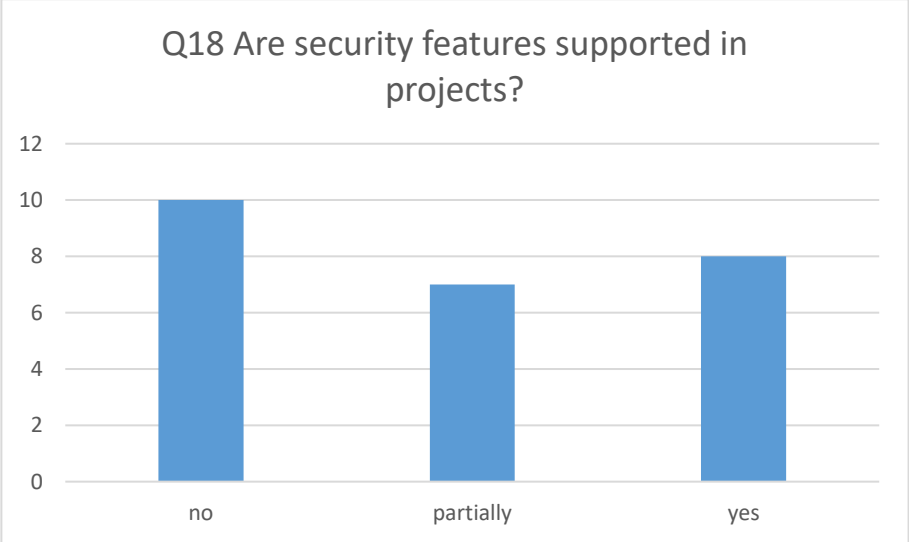
Data description:
It can be clearly observed that both cold and hot data. Cold data is typically less expensive to store, since it is often stored on slower, less expensive storage devices like tape or hard drives. This makes it an ideal solution for storing historical data, backups, and archives, which may not need to be accessed frequently, but need to be stored for compliance or legal reasons. Hot data, on the other hand, is typically stored on faster, more expensive storage devices. This type of data is frequently accessed by applications, users, or services, and needs to be quickly available in order to support real-time operations, transactions, or analytics. In most organizations, both hot and cold data are necessary for business operations.

Discussion:

- Both cold and hot data are common equally because they serve different purposes and are needed at different times. Cold data refers to data that is accessed infrequently or not at all, while hot data refers to data that is frequently accessed or actively being used.

3.15. Are security features supported in projects?

The next question asks whether security features are supported in projects. The alternatives are yes, no or partially**.**



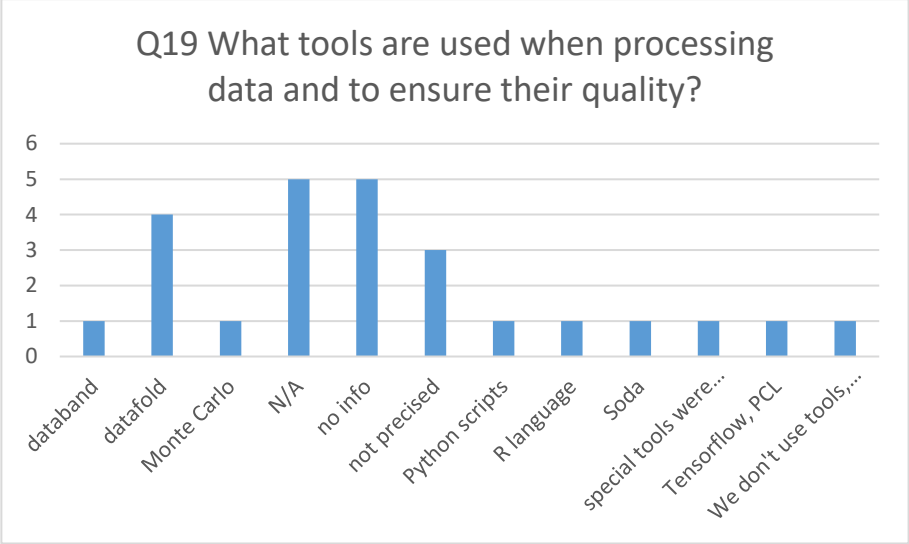
Data description:
The results indicate that a bit more often security features are not supported than vice versa. A possible explanation could be that implementing robust security features can take time and resources and may require significant investments in hardware, software, and personnel. In some cases, organizations may prioritize other features or functionalities over security due to budget or time constraints. Another probable reason may be the lack of enough expertise because security can be a complex and specialized field and not all developers or organizations may have the resources or knowledge to properly address security concerns. However, the differences are small, so it can't be generalized that security features are not supported, only in some cases.

Discussion:

- Security features are not supported in most projects.
- It is important for organizations and developers to prioritize security and take steps to implement robust security measures to protect their users and their data.

3.16. What tools are used when processing data and to ensure their quality?

That questions suggests different options, regarding tools when processing data. The options are Talend, Toro, Soda, Datafold, Databand, Precisely, Monte Carlo or other. The distribution of the results are shown in the chart below.



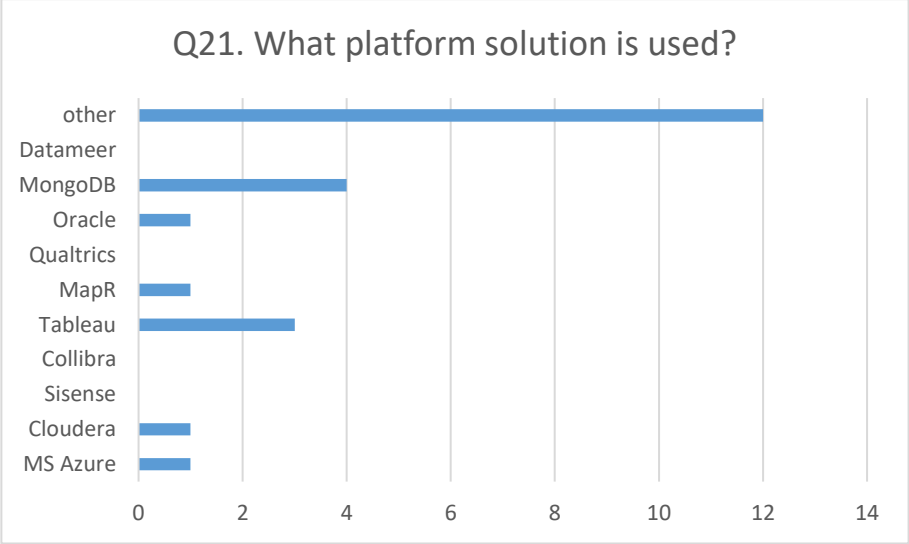
Data description:
The results are predominantly "N/A" or "no information" which suggests that in the studies analyzed, the scientists have not mentioned the tools which they used when processing data. Datafold takes the third position. It is becoming more widely used and is considered a valuable tool for data processing workflows.

Discussion:

- Tools, which are used when processing data, are not often mentioned.
- Datafold is a relatively new tool for data processing that has gained popularity among data engineers and data scientists.

3.17. What platform solution is used?

The next question reveals what platform solution is used. The possible answers are MS Azure, Cloudera, Sisense, Collibra, Tableau, MapR, Qualtrics, Oracle, MongoDB, Datameer or other which is entered by the respondent. The results are shown in the chart.



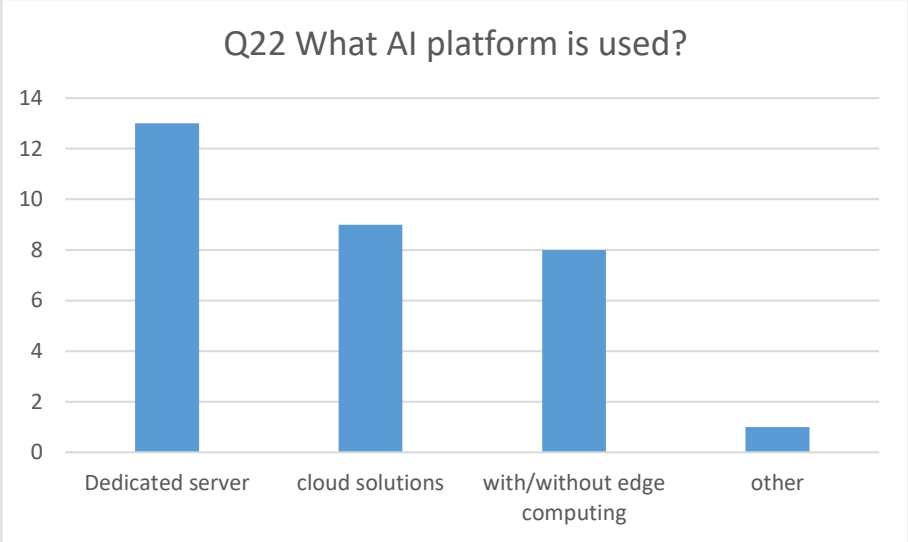
Data description:
The platform solutions used in the projects are predominantly other. They include R studio, NVIDIA CUDA, Clara, CUDA-X, TensorFlow/TensorRT, Anaconda, Google Colab, Apache Spark, Apache Flink, PySpark, AWS, Bonseyes, specialized Radiology Information System, Amazon Web Services S3 Data Lake, FedML, OpenMMLab, gRPC.

Discussion:

- The platform solutions which are suggested as answers are not so popular. MongoDB and Tableau among them take the second and third position respectively.
- Other platforms, such as Anaconda, Apache, NVIDIA, R studio, TensorFlow, etc. are used in more cases.

3.18. What AI platform type is used (e.g. server-based, cloud solutions, with/without edge computing support or other)?

The following question is directed to the type of AI platform. The possible alternatives are dedicated server, cloud solutions, with/without edge computing support or other types, entered by the respondents. The results are presented in the chart.



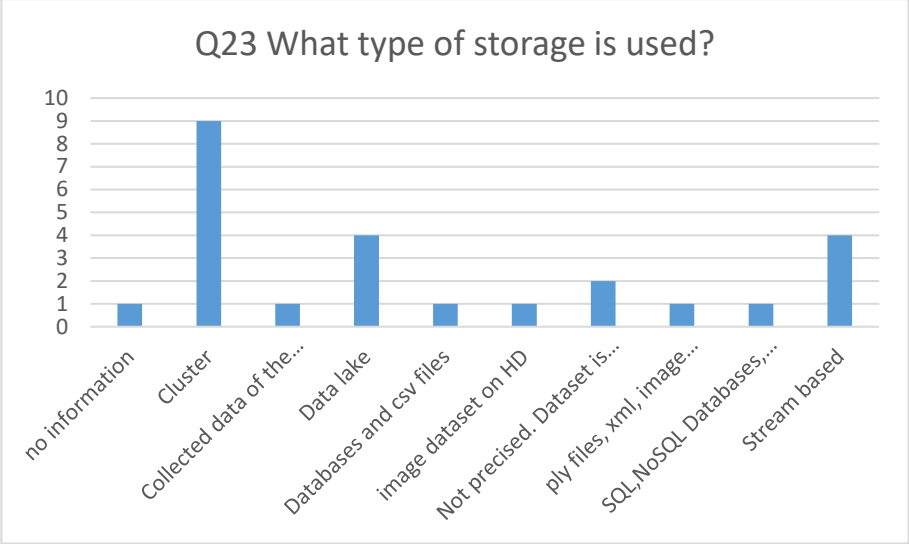
Data description:
The results indicate that the most used AI platform type is the dedicated server. Dedicated servers are often used as AI platform types because they offer a number of advantages that make them well-suited for running AI workloads. Those servers can be customized, their performance is high and are often considered more secure than shared hosting solutions. In addition, dedicated servers can be scaled up or down and offer more control over the server environment than shared hosting solutions.

Discussion:

- The high performance, customizability, security, scalability, and control offered by dedicated servers make them a popular choice for running AI workloads.

3.19. What type of storage is used?

The next questions gives details about the type of storage. The options suggested are cluster, stream based, data lake or other. The distribution of the responses is shown in the chart.



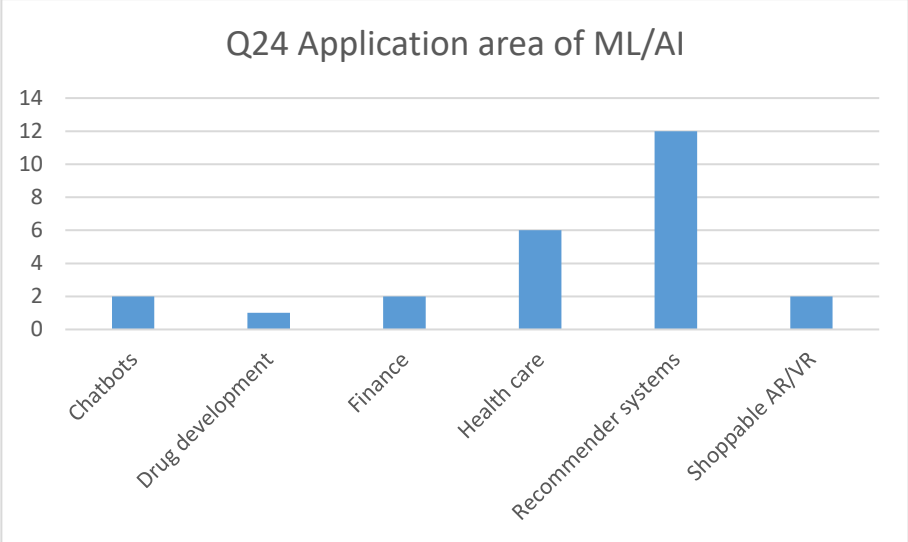
Data description:
The results indicate that cluster is the most popular type of storage among the researchers. The possible explanation could be that clusters are designed to provide high availability of data. Clusters are highly scalable and fault-tolerant. Clusters can provide high levels of performance and can be a cost-effective storage solution, as they can be built using commodity hardware.

Discussion:

- Clusters are popular as a type of storage because they offer high availability, scalability, fault tolerance, performance, and cost-effectiveness. These benefits make them an ideal choice for businesses and applications that require reliable and scalable storage solutions.

3.20. Application area of ML/AI?

The following questions reveals the application areas of machine learning/artificial intelligence. The question suggests eight options – recommender systems, chatbots, A/B tests, shoppable augmented reality-AR/VR, health care, drug development, finance, cybersecurity. The results are presented in the chart below.



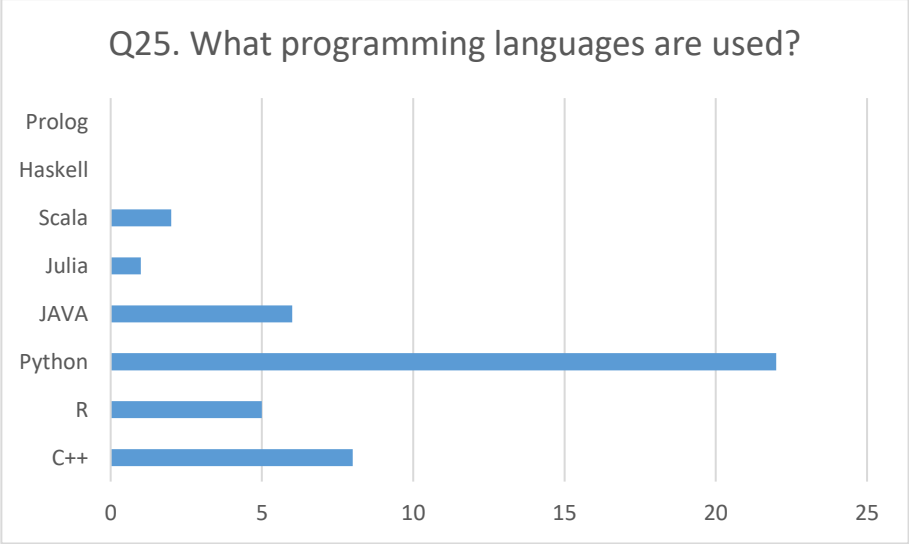
Data description:
The data from the projects reviewed show that the most common area, where ML and AI are applied, are the recommender systems. Recommender systems are designed to provide personalized recommendations to users based on their preferences. Recommender systems typically work with Big Data, such as user behavior and product information. Many recommender systems are required to provide recommendations in real-time, such as for e-commerce sites or streaming platforms. Recommender systems can provide significant business benefits, such as increased sales, customer engagement, and loyalty.

Discussion:

- Recommender systems are a common area of application of artificial intelligence because they require analyzing complex data sets to provide personalized recommendations in real-time. AI algorithms can help to automate and improve this process, which can lead to better business outcomes and improved user experiences.

3.21. What programming languages are used?

This question points the programming languages which are used in AI problems. The results are presented in the chart.



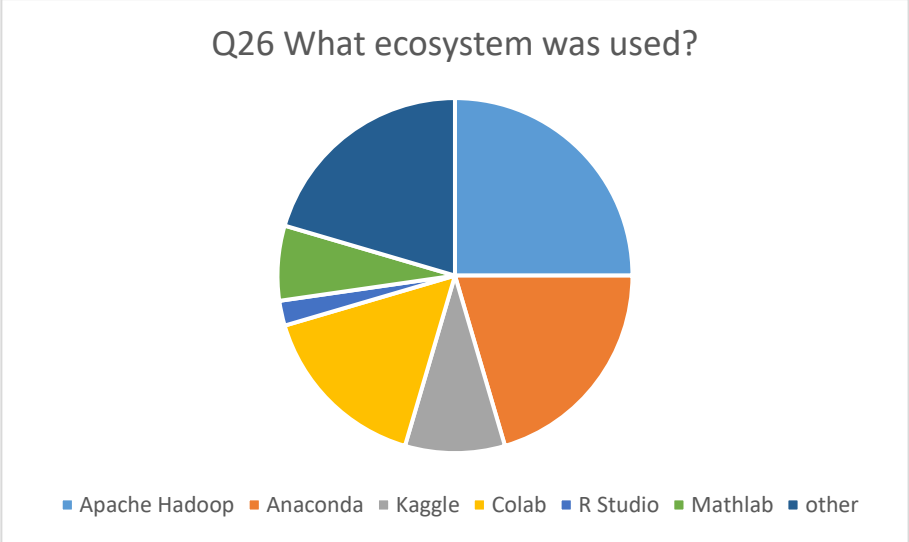
Data description:
The results clearly show that the most used programming language is Python. Python has a large and active community of developers who have created numerous libraries and frameworks for AI development. Python is a flexible language that can be used for a wide range of AI tasks, including data processing, machine learning, and natural language processing. In addition, Python is compatible with a wide range of platforms and systems, including Windows, Mac, and Linux. While Python is not the fastest programming language, it is fast enough for most AI tasks.

Discussion:

- Python is a popular programming language for artificial intelligence because of its ease of use, large community, flexibility, compatibility, and performance. These factors make it an ideal choice for building and deploying AI applications.

3.22. What ecosystem was used?

The next question gives information about the ecosystem which was used. These are the possible alternatives- Apache Hadoop, Anaconda, Kaggle, Colab, R studio, Mathlab or other. The distribution of the responses is presented in the chart below.



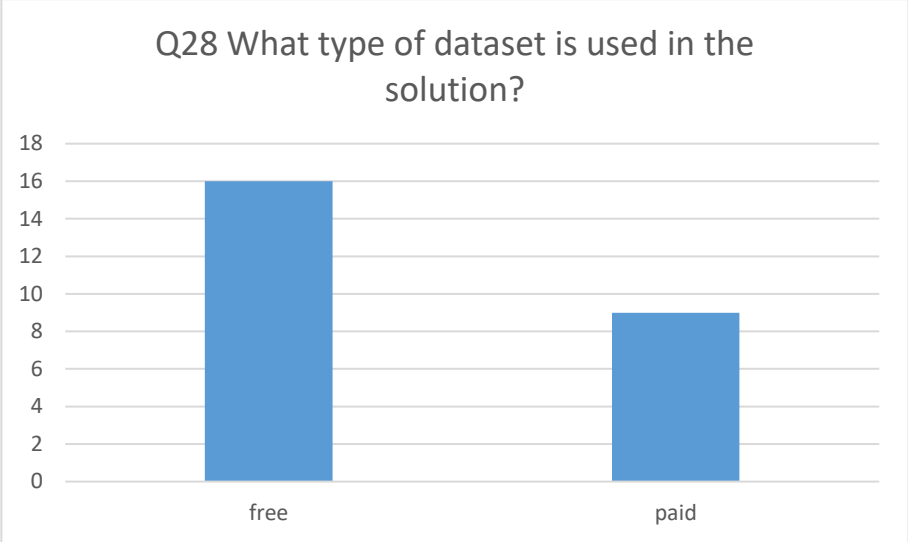
Data description:
The results indicate that Apache Hadoop was used the most, with a slight predominance over Anaconda and other systems. The option "other"includes ecosystems, such as NVIDIA CUDA, TensorFlow, local Python IDE, Apache Spark, DataBricks.

Discussion:

- The Apache Hadoop ecosystem is widely used for big data processing and analysis because of its scalability, distributed processing, fault tolerance, open-source nature, large ecosystem, and industry adoption. These factors make it an ideal solution for organizations that need to process and analyze large amounts of data.
- Anaconda and other ecosystems related to Python language mostly, are other popular ecosystems, too. Anaconda is a distribution of the Python and R programming languages, along with a collection of open-source libraries, tools, and frameworks for data science and scientific computing.

3.23. What type of dataset is used in the solution?

The next question provides information about the type of the dataset – whether it is free or paid. The results are presented in the pie chart.



Data description:
The results show a predominance of free datasets which are used in the solutions. Free datasets are more accessible to a wider range of users, including students, researchers, and developers. Many free datasets are open data, which means that they are freely available for anyone to use, modify, and distribute. Paid datasets often come with restrictions on how they can be used, which can limit their usefulness for certain types of analysis or applications.

Discussion:

- The availability of free datasets has contributed to the growth and development of the data science and machine learning fields by promoting accessibility, openness, collaboration, and innovation.

CONCLUSIONS

So, the comprehensive analysis of the collected data provides valuable insights into the current landscape of artificial intelligence (AI) and machine learning (ML) projects undertaken by various research institutions. The key findings can be summarized as follows.
The data collection involved 25 questionnaires from five partner institutions. Most training in AI takes place in Serbia, followed by Bulgaria. The survey underscores a need for increased training opportunities, particularly in European Union countries, to foster good practices in AI.

Deep Machine Learning (ML) is the most commonly used model, demonstrating its dominance in various AI applications. Other ML models, such as SciMl, are less prevalent among the surveyed projects.

Convolutional Neural Networks (CNNs) are the most commonly developed models, reflecting their versatility in tasks like image classification, object detection, and medical image analysis. Gated Recurrent Unit (GRU) was not developed in the surveyed projects.

Data volumes between 1 GB and 1 TB are most commonly processed, aligning with the capabilities of modern computers. Extremely large datasets (over 1 TB) or very small datasets (less than 1 GB) are less prevalent, indicating practical considerations and resource limitations.

AI solutions are implemented across diverse fields, including healthcare, finance, smart cities, and Industry 4.0. Each respondent explored specific areas, providing a broad perspective on AI applications.

TensorFlow emerges as the most widely used AI library, followed by Keras and Scikit-learn. The choice of library depends on project-specific needs, developer preferences, and expertise levels.

Results vary widely across projects, encompassing applications like human body detection, COVID-related analytics, and object detection. Each respondent uncovered unique outcomes tailored to their specific project.

Permissive licenses are the most prevalent, offering flexibility and collaboration opportunities. Proprietary licenses are the second most common, while copyleft licenses impose more restrictions.

Databases are the primary source of data, providing structured and efficient storage.

Applications and forms are less common due to scattered data and inconsistent quality.

Texts and pictures dominate as data characteristics, reflecting their ease of understanding, generation, and versatility.

Various methods, including expert decisions, Python formatting, and classification, are employed to process data and ensure quality.

NoSQL databases are the most commonly used tools for data storage, offering flexibility and scalability.

Both cold and hot data are prevalent, meeting different needs and usage frequencies.

Security features are not consistently supported across projects, with resource constraints and expertise cited as potential reasons.

Mentioned tools for data processing are not consistently provided, with Datafold emerging as one of the tools.

Other platforms, including R studio, NVIDIA CUDA, and TensorFlow, dominate over suggested options like MongoDB and Tableau.

Dedicated servers are the most commonly used AI platform type, offering high performance, customizability, and security.

Clusters are the preferred type of storage, providing high availability, scalability, and fault tolerance.

Recommender systems are the most commonly applied areas of ML/AI, leveraging personalized recommendations.

Python is the overwhelmingly favored programming language for AI projects due to its community support, flexibility, and compatibility.

Apache Hadoop is the most used ecosystem, followed by Anaconda and other Python-related ecosystems.
Free datasets dominate, fostering accessibility, openness, collaboration, and innovation in the AI and ML fields.
In conclusion, the diverse range of findings highlights the dynamic and evolving nature of AI and ML projects, emphasizing the importance of tailored solutions and continuous adaptation to emerging technologies and methodologies. The identified trends and patterns provide a foundation for future research and development in the rapidly advancing field of artificial intelligence.

REFERENCES

1. https://towardsdatascience.com/why-deep-learning-is-needed-over-traditional-machine-learning-1b6a99177063
2. https://www.ibm.com/topics/convolutional-neural-networks
3. https://hub.packtpub.com/tensorflow-always-tops-machine-learning-artificial-intelligence-tool-surveys/
4. https://blog.ipleaders.in/permissive-license-copyleft-possible-distinctions/
5. https://www.techtarget.com/searchdatamanagement/definition/database
6. https://www.simplilearn.com/rise-of-nosql-and-why-it-should-matter-to-you-article
7. https://www.dataversity.net/cold-vs-hot-data-storage-whats-the-difference/
8. https://datalogistics.lt/en/dedicated-servers-are-an-increasingly-popular-hosting-service/
9. https://www.techtarget.com/searchstorage/magazineContent/The-benefits-of-clustered-storage
10. Roy, D., Dutta, M. A systematic review and research perspective on recommender systems. J Big Data 9, 59 (2022). https://doi.org/10.1186/s40537-022-00592-5
11. https://www.pulumi.com/why-is-python-so-popular/
12. https://www.projectpro.io/article/apache-hadoop-turns-10-the-rise-and-glory-of-hadoop/211
13. https://towardsdatascience.com/an-overview-of-the-anaconda-distribution-9479ff1859e6
14. Sakshi Indolia, Anil Kumar Goswami, S.P. Mishra, Pooja Asopa,Conceptual Understanding of Convolutional Neural Network- A Deep Learning Approach, Procedia Computer Science, Volume 132, 2018, Pages 679-688,ISSN 1877-0509, https://doi.org/10.1016/j.procs.2018.05.069.