Co-funded by the European Union

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

01.09.2022 – 31.08.2024

Badanie 7: Gromadzenie specyfikacji IT dobrych praktyk w zakresie sztucznej inteligencji:
analiza dla WP2

Co-funded by the European Union

Powstanie tego dokumentu było możliwe dzięki wsparciu projektu ERASMUS+: The Future is in Applied Artificial Intelligence (2022-1-PL01-KA220-HED-000088359)

Data
15.03.2023

Miejsca rozwoju wyniku
Uniwersytet Bielsko-Bialski, Bielsko-Biała, Polska
Uniwersytet Bibliotekoznawstwa i Technologii Informacyjnych, Sofia, Bułgaria
Uniwersytet w Niszu, Serbia
Uniwersytet św. Cyryla i Metodego w Trnawie, Słowacja
Uniwersytet Czarnogóry, Czarnogóra

Streszczenie: W pracy przedstawiono opracowanie specyfikacji dobrych praktyk w zakresie stosowanej sztucznej inteligencji (AAI). Analiza 25 kwestionariuszy z pięciu instytucji partnerskich ujawniła kluczowe spostrzeżenia na temat obecnego stanu projektów sztucznej inteligencji (AI) i uczenia maszynowego (ML). Szkolenia przeprowadzone w Serbii i Bułgarii sygnalizowały potrzebę rozszerzenia możliwości w krajach UE. W wyniku przeprowadzonych badań uzyskaliśmy, że Deep ML przeważa, szczególnie w konwolucyjnych sieciach neuronowych, podczas gdy Gated Recurrent Unit jest mniej powszechny. Typowe są ilości danych od 1 GB do 1 TB, co odzwierciedla ograniczenia praktyczne. Aplikacje AI obejmują różne dziedziny, a TensorFlow jest liderem w dziedzinie bibliotek. Najbardziej rozpowszechnione są licencje permisywne, bazy danych są podstawowymi źródłami danych, a teksty/obrazy dominują w charakterystyce danych. Bazy danych NoSQL są preferowane do przechowywania. Funkcje bezpieczeństwa i narzędzia do przetwarzania danych są różne. Serwery dedykowane i klastry są szeroko stosowane, systemy rekomendacji są widoczne, preferowanym językiem jest Python, a Apache Hadoop dominuje w ekosystemach. Bezpłatne zestawy danych sprzyjają dostępności. Ogólnie rzecz biorąc, odkrycia podkreślają dynamiczny charakter projektów AI/ML, zapewniając podstawę dla przyszłych badań w szybko rozwijającej się dziedzinie.

Słowa kluczowe: sztuczna inteligencja stosowana, uczenie maszynowe, dobra praktyka, konwolucyjne sieci neuronowe

1. Wprowadzenie

Sztuczna inteligencja (AI) i uczenie maszynowe (ML) stały się integralnymi elementami najnowocześniejszych badań, napędzając innowacje i transformacyjne rozwiązania w różnych branżach. W kompleksowej analizie zagłębiamy się w krajobraz projektów AI i ML podejmowanych przez różne instytucje badawcze. To badanie, oparte na danych zebranych z 25 kwestionariuszy w pięciu instytucjach partnerskich, ujawnia krytyczne spostrzeżenia, które kształtują obecny stan inicjatyw w zakresie sztucznej inteligencji i uczenia maszynowego.

Odkrycia ujawniają bogaty gobelin trendów, wyzwań i możliwości w dziedzinie sztucznej inteligencji. Od dynamiki szkolenia w Serbii i Bułgarii po powszechne wykorzystanie modeli głębokiego uczenia maszynowego, badanie nakreśla zniuansowany obraz krajobrazu sztucznej inteligencji. W miarę jak przechodzimy przez kluczowe wyniki, staje się oczywiste, że zastosowania sztucznej inteligencji obejmują wiele dziedzin, w tym opiekę zdrowotną, finanse, inteligentne miasta i Przemysł 4.0.

TensorFlow wyłania się jako lider wśród bibliotek AI, podkreślając znaczenie wyboru narzędzi dostosowanych do konkretnych potrzeb projektu. Rozpowszechnienie konwolucyjnych sieci neuronowych (CNN) w różnych zastosowaniach, wraz z niedoborem niektórych modeli uczenia maszynowego, takich jak SciMl, odzwierciedla zmieniające się preferencje i wyzwania w tej dziedzinie.

Zagadnienia dotyczące danych, od ilości przetwarzania od 1 GB do 1 TB po dominację bezpłatnych zestawów danych, zapewniają wgląd w praktyczne aspekty i dostępność kształtujące przedsięwzięcia związane ze sztuczną inteligencją. Znaczenie licencji permisywnych, dominujące wykorzystanie baz danych NoSQL i zniuansowany charakter implementacji funkcji bezpieczeństwa dodatkowo charakteryzują skomplikowany krajobraz projektów AI.

W miarę przechodzenia przez szczegóły charakterystyki danych, metodologii przetwarzania i różnych osiągniętych wyników, staje się oczywiste, że sztuczna inteligencja nie jest dziedziną uniwersalną. Dynamiczny charakter projektów AI wymaga rozwiązań dostosowanych do indywidualnych potrzeb i ciągłego dostosowywania się do pojawiających się technologii.

Celem badania jest fundamentalna eksploracja dynamicznej i ewoluującej dziedziny sztucznej inteligencji. Zidentyfikowane trendy i wzorce działają jako drogowskazy dla przyszłych badań i rozwoju, podkreślając potrzebę zdolności adaptacyjnych i innowacji w szybko rozwijającej się dziedzinie sztucznej inteligencji.

2. Gromadzenie i analiza danych

Dane zostały pozyskane przez naukowców z pięciu instytucji partnerskich. W niniejszym badaniu uzyskano 8 kwestionariuszy z UBB, 6 kwestionariuszy z ULSIT, 5 z UNi, 5 z UoM i 1 z USCM.
Łącznie 11 badaczy zebrało 25 kwestionariuszy.

3. Wyniki

3.1. Kraj, w którym odbywa się szkolenie

Pierwsze pytanie ankiety, po danych dotyczących nazwy organizacji i badacza, dotyczy miejsca, w którym odbywa się szkolenie. Wyniki przedstawiono poniżej.

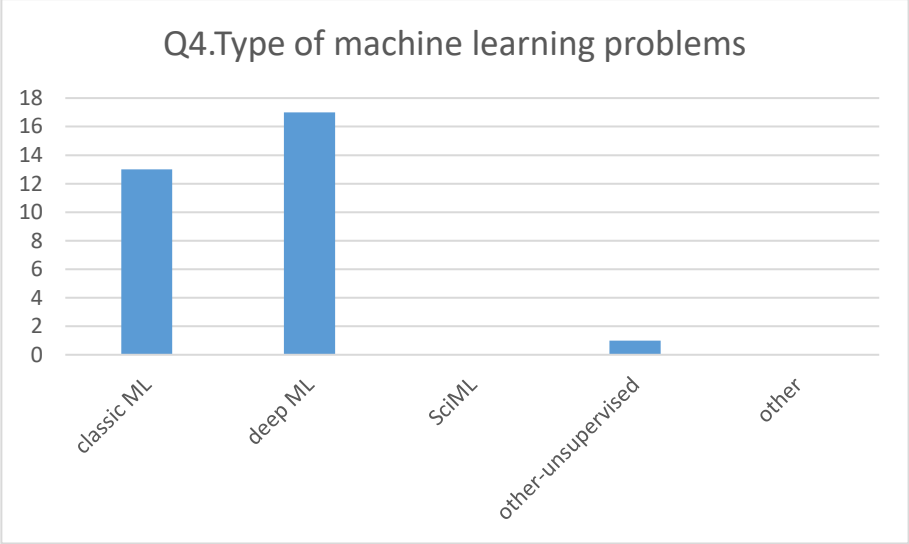
Opis danych:
Z danych badawczych wynika, że szkolenia odbywają się głównie w Serbii, a następnie w Bułgarii. Jeden z respondentów odpowiedział, że używał Coursera lub O'Reilly tylko do szkoleń. W pozostałych krajach objętych badaniem odbywa się od jednego do trzech szkoleń.

Dyskusja:
Z badania wynika, że istnieje potrzeba większej liczby szkoleń dla specjalistów IT, aby można było zaobserwować więcej dobrych praktyk w zakresie AI. **Główne wnioski:**

- Istnieje potrzeba większej liczby szkoleń w krajach UE.
- Najwięcej szkoleń odbywa się w Serbii i Bułgarii.

3.2. Rodzaj problemów związanych z uczeniem maszynowym

Kolejne pytanie ujawnia rodzaj modeli uczenia maszynowego, które zostały wykorzystane.



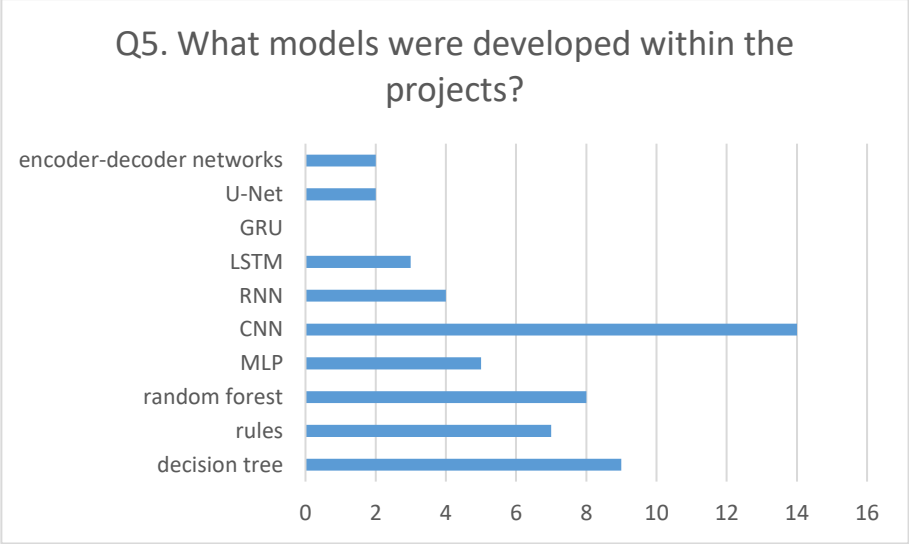
Opis danych:
Dane pokazują, że najczęściej używanym modelem uczenia maszynowego jest głębokie uczenie maszynowe, a następnie klasyczne uczenie maszynowe. W ciągu ostatnich lat głębokie uczenie stało się głównym narzędziem do rozwiązywania różnych problemów związanych ze sztuczną inteligencją. Wykazał lepszą wydajność w różnych zadaniach, w tym przetwarzaniu języka naturalnego, wizji, wirtualnych asystentach, chatbotach, opiece zdrowotnej itp.

Dyskusja:

- Najczęściej używanym modelem, na podstawie ankiety, jest deep ML.
- SciMl lub inne modele uczenia maszynowego nie są powszechne wśród recenzowanych projektów.

3.3. Jakie modele zostały opracowane (zbadane) w ramach projektów?

Poniższe pytanie dotyczy tego, jakie modele były badane w szczególności. Wyniki przedstawiono na poniższym wykresie.

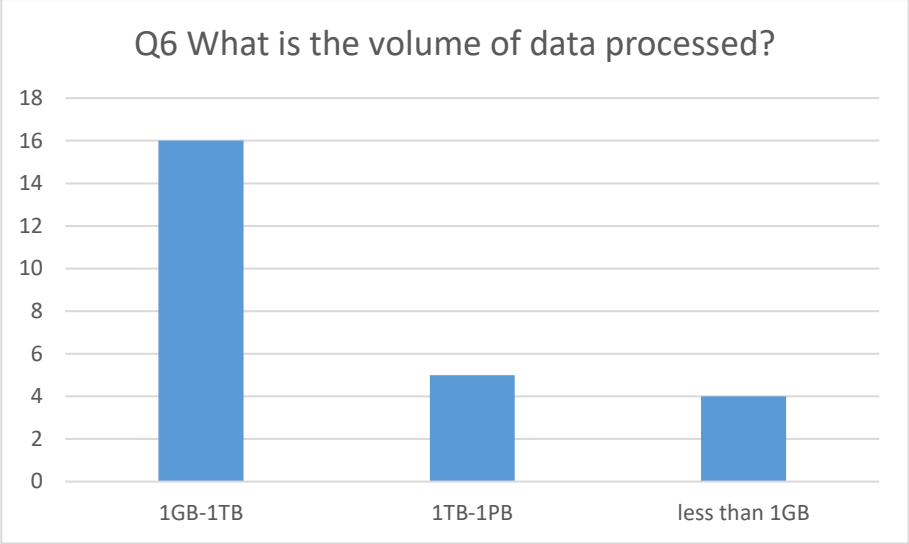


Opis danych:
Zgodnie z wynikami, modelami najczęściej wykorzystywanymi w projektach były konwolucyjne sieci neuronowe. Jest to zrozumiałe, ponieważ modele te mają różnorodne zastosowania – klasyfikacja obrazów, wykrywanie obiektów, rozpoznawanie twarzy, analiza obrazu medycznego, NLP, pojazdy autonomiczne czy technologia. Sieci CNN są szczególnie przydatne do zadań związanych z rozpoznawaniem obrazów, ponieważ mogą automatycznie nauczyć się wykrywać złożone elementy, takie jak krawędzie, narożniki i tekstury. Drzewo decyzyjne, wraz z lasem losowym, to inne modele, które zostały opracowane dla różnych problemów związanych ze sztuczną inteligencją. Wiele zastosowań drzew decyzyjnych i lasów losowych jest szeroko stosowanych w branżach takich jak finanse, opieka zdrowotna, produkcja i monitorowanie środowiska. Drzewa decyzyjne i lasy losowe są popularnymi algorytmami uczenia maszynowego, ponieważ są łatwe w użyciu, zrozumieniu i interpretacji, mogą obsługiwać zarówno dane liczbowe, jak i jakościowe oraz mogą przetwarzać duże i złożone zbiory danych.

Dyskusja:
- Konwolucyjne sieci neuronowe są najczęściej spotykanymi modelami w projektach.
- GRU nie powstało w ramach projektów.

3.4. Jakie modele zostały opracowane (zbadane) w ramach projektów?

Kolejne pytanie wskazuje na ilość danych, które zostały przetworzone. Możliwe odpowiedzi to mniej niż 1 GB, 1 GB – 1 TB, 1 TB – 1 PB i ponad 1 PB.

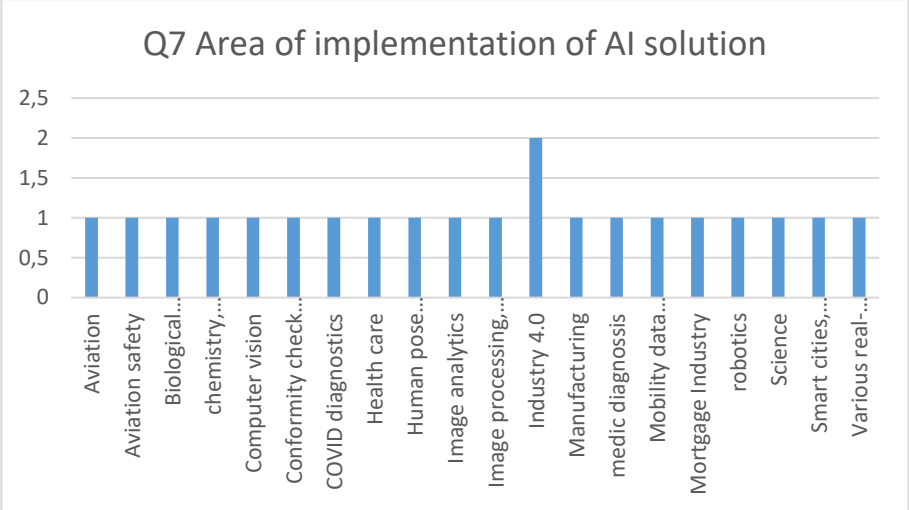


Opis danych:
Jak widać na wykresie, ilość przetwarzanych danych mieściła się głównie w przedziale od 1 GB do 1 TB. Wolumen ten reprezentuje stosunkowo duże ilości danych, które są powszechnie spotykane w wielu różnych typach projektów, ale nadal są zarządzalne pod względem możliwości przechowywania i przetwarzania najnowocześniejszych komputerów. Ilość przetwarzanych danych powyżej 1 TB nie jest tak powszechna, prawdopodobnie ze względu na dostępność i koszty. Zarządzanie danymi i ich przetwarzanie w skali petabajtów wymaga specjalistycznej infrastruktury i zasobów, które nie są łatwo dostępne dla większości organizacji. Przechowywanie i przetwarzanie dużych ilości danych może być kosztowne, zarówno pod względem kosztów sprzętowych, jak i operacyjnych. Z drugiej strony, objętość mniejsza niż 1 GB również nie jest powszechna, ponieważ uważa się, że jest wystarczająco mała, aby można ją było wydajnie przetwarzać na jednym komputerze, bez potrzeby stosowania specjalistycznego sprzętu lub rozproszonych systemów obliczeniowych.

Dyskusja:
- Najczęściej wykorzystywana jest ilość przetwarzanych danych o pojemności od 1 GB do 1 TB.
- Wolumin większy niż 1 TB lub mniejszy niż 1 GB nie ma zastosowania.

3.5. Jaki jest obszar wdrożenia rozwiązania AI?

Kolejne pytanie skierowane jest do obszarów wdrażania rozwiązań AI. Odpowiedzi zawierają wkład własny respondentów, w oparciu o każdy przypadek. Wyniki przedstawiono graficznie poniżej.

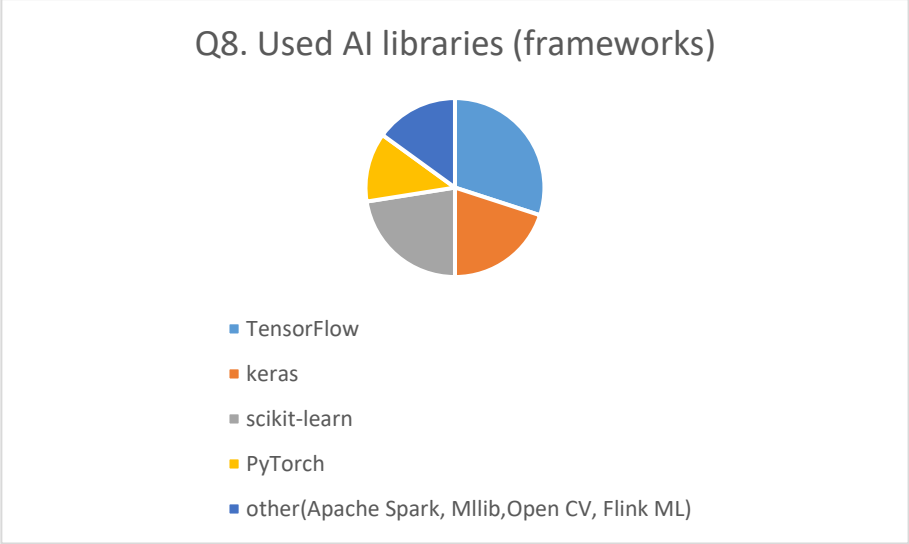


Opis danych:
Należy zaznaczyć, że naukowcy poszukiwali projektów z różnych dziedzin, którymi są: rolnictwo, sztuczna inteligencja w medycynie, chirurgia, zarządzanie ruchem lotniczym, lotnictwo, bezpieczeństwo lotnicze, analiza sekwencji biologicznych, chemia, robotyka, zdrowie, wizja komputerowa, kontrola zgodności w przemyśle lotniczym, diagnostyka COVID, opieka zdrowotna, estymacja pozycji człowieka, analiza obrazu, przetwarzanie obrazu, prognozowanie cen, przemysł 4.0, produkcja, diagnostyka medyczna, nauka i analiza danych dotyczących mobilności, branża kredytów hipotecznych, nauka, inteligentne miasta, monitorowanie ruchu, różne przypadki ze świata rzeczywistego. Każdy z respondentów zbadał jeden lub dwa z tych obszarów wdrażania rozwiązań AI.

Dyskusja:
- Każdy respondent został przydzielony do zbadania określonego obszaru, tak aby można było przejrzeć jak najwięcej obszarów.

3.6. Wykorzystane biblioteki (frameworki) AI.

W pytaniu 8 zasugerowano różne biblioteki sztucznej inteligencji, które są zwykle wykorzystywane w dziedzinie sztucznej inteligencji. Są one pokazane na poniższym wykresie kołowym.

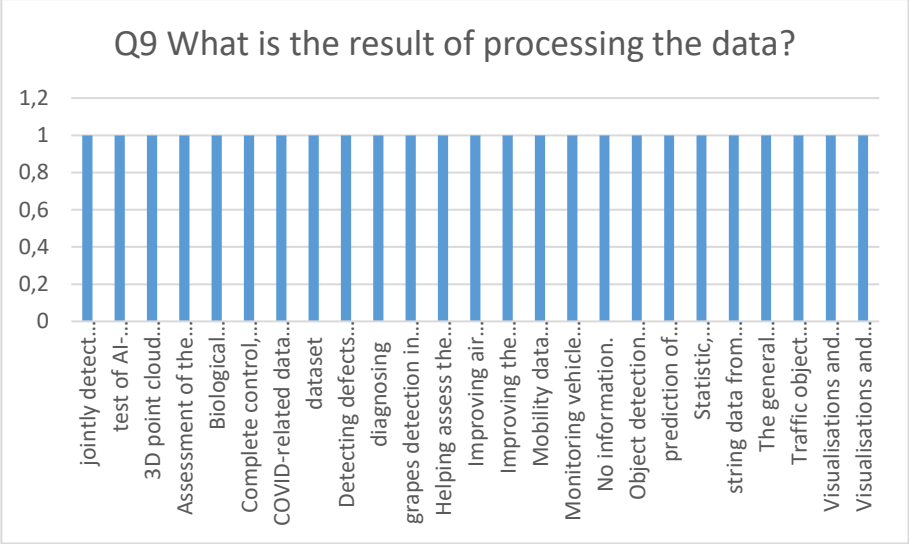


Opis danych:
Wyniki wskazują, że najczęściej używaną biblioteką AI jest TensorFlow. TensorFlow to jedna z najpopularniejszych i najczęściej używanych bibliotek typu open source do tworzenia i wdrażania modeli uczenia maszynowego i głębokiego uczenia. Popularność tej biblioteki można również przypisać jej licznej i aktywnej społeczności, która zapewnia wsparcie i przyczynia się do jej rozwoju. Inne popularne biblioteki AI to Keras i Scikit-learn. Inne biblioteki używane w problemach sztucznej inteligencji obejmują PyTorch, Apache Spark, MLlib, Open CV i Flink ML.

Dyskusja:
- Wybór biblioteki zależy od różnych czynników, w tym specyficznych potrzeb projektu, osobistych preferencji i poziomu wiedzy programisty.

3.7. Jaki jest rezultat przetwarzania danych?

Pytanie to zawiera otwarte odpowiedzi respondentów, w zależności od każdego przypadku. W związku z tym dla każdego projektu uzyskano dwadzieścia pięć różnych wyników.



Opis danych:
W pierwszym projekcie rezultatem było wspólne wykrywanie kluczowych punktów ludzkiego ciała, twarzy i stóp. W drugim projekcie rezultatem był test metod AI-TWILIGHT w branży motoryzacyjnej, ogrodniczej i oświetlenia ulicznego. Trzecim rezultatem była klasyfikacja i segmentacja chmur punktów 3D. Czwartym rezultatem była ocena prawdopodobieństwa i ciężkości zakażenia Covid-19. Piątym wynikiem była klasyfikacja sekwencji biologicznych. Szóstym rezultatem była pełna kontrola, pełna artykulacja i inteligentne sprzężenie zwrotne. Siódmym rezultatem była analiza danych związanych z COVID. Ósmym rezultatem było utworzenie zbioru danych. Dziewiąty wynik dotyczył wykrywania wad produktów w produkcji. Dziesiąty wynik został skierowany do wdrożenia w diagnostyce. Jedenastym rezultatem było wykrycie winogron w winnicy. Dwunastym rezultatem była pomoc w ocenie stanu maszyn i ogólnego wyposażenia w fabryce. Trzynastym rezultatem było usprawnienie zarządzania ruchem lotniczym poprzez współpracę w zakresie uczenia maszynowego na prywatnych zbiorach danych. Czternastym rezultatem było usprawnienie procesu decyzyjnego w scenariuszach odejścia na drugi krąg, co ma ogromne znaczenie dla bezpieczeństwa zarówno linii lotniczych, jak i instytucji zapewniających służby żeglugi powietrznej w ATM. Piętnastym rezultatem była analiza danych dotyczących mobilności. Szesnastym rezultatem był monitoring pojazdu, który porusza się po miastach, zliczając samochody ze zdjęć pozyskanych z inteligentnych kamer. Siedemnastym rezultatem było wykrywanie i rozpoznawanie obiektów. Osiemnastym wynikiem była predykcja aktywności związków chemicznych, automatyczne sterowanie robotami, analiza ruchów człowieka z czujników ubieralnych. Dziewiętnastym wynikiem były statystyki, przepisywanie leków i wyznaczanie metod leczenia. Dwudziestym rezultatem było uzyskanie danych z obrazów i obiektu z obrazów. Dwudziestym pierwszym rezultatem była próba opracowania prototypu globalnego systemu monitorowania i wczesnego ostrzegania przed wieloma zagrożeniami. Dwudziestym drugim wynikiem było wykrywanie i rozpoznawanie obiektów drogowych. Dwudziesty trzeci wynik prezentował wizualizacje i informacje

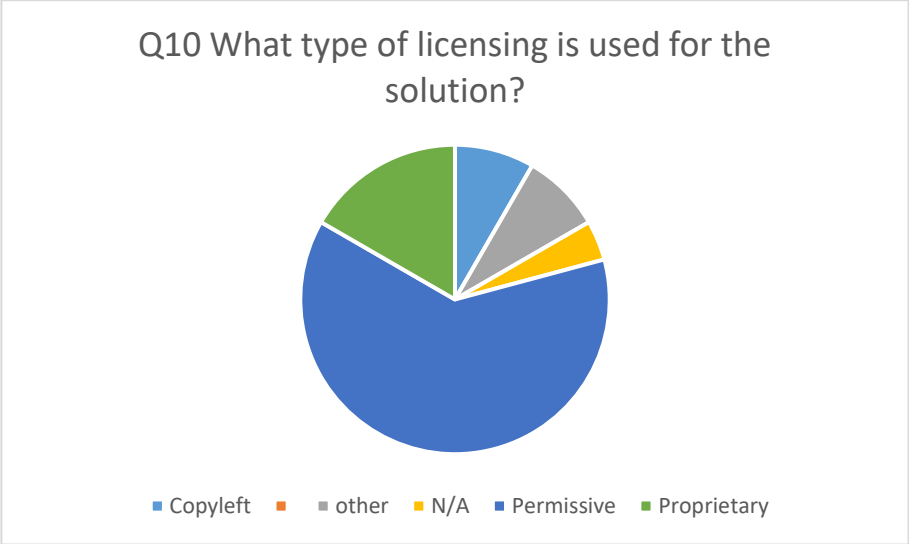
o przyszłych cenach. W dwudziestym czwartym wyniku zaprezentowano wizualizacje i modele ML. Był jeden projekt, w którym nie udokumentowano żadnych wyników.

Dyskusja:

- Każdy respondent musiał znaleźć wyniki po przetworzeniu danych, w zależności od każdego projektu. W ten sposób odkryto różne wyniki w dziedzinach sztucznej inteligencji.

3.8. Jaki rodzaj licencjonowania jest używany w rozwiązaniu?

Poniższe pytanie dotyczy tego, jaki rodzaj licencjonowania jest używany w rozwiązaniu i obejmuje cztery alternatywy – permisywne (BSD, MIT), copyleft (GPL, LGPL), własnościowe (Bespoke, Commercial) i inne. Poniższy wykres przedstawia rozkład tych typów licencjonowania.



Opis danych:

Wyniki pokazują, że najczęściej stosowanym typem licencji jest licencja permisywna (w 15 projektach), następnie licencja własnościowa (w 4 projektach), copyleft (w 2 projektach). W jednym z projektów wykorzystano dostępną do niekomercyjnego użytku oryginalną licencję OpenPose . Licencje permisywne są popularne, ponieważ oferują wysoki stopień swobody użytkownikom i programistom. Tego typu licencje zazwyczaj pozwalają użytkownikom na modyfikowanie i rozpowszechnianie oprogramowania bez konieczności wprowadzania jakichkolwiek zmian lub ulepszeń w ramach tej samej licencji. Jest to przeciwieństwo licencji typu copyleft, które wymagają, aby wszelkie dzieła pochodne były licencjonowane na tych samych warunkach, co oryginał. Licencje permisywne są często wybierane przez osoby i organizacje, które chcą zachęcać do współpracy i innowacji, a jednocześnie pozwalają na maksymalną elastyczność w sposobie korzystania i dystrybucji oprogramowania. Licencje własnościowe zajmują drugie miejsce, chociaż występują tylko w czterech

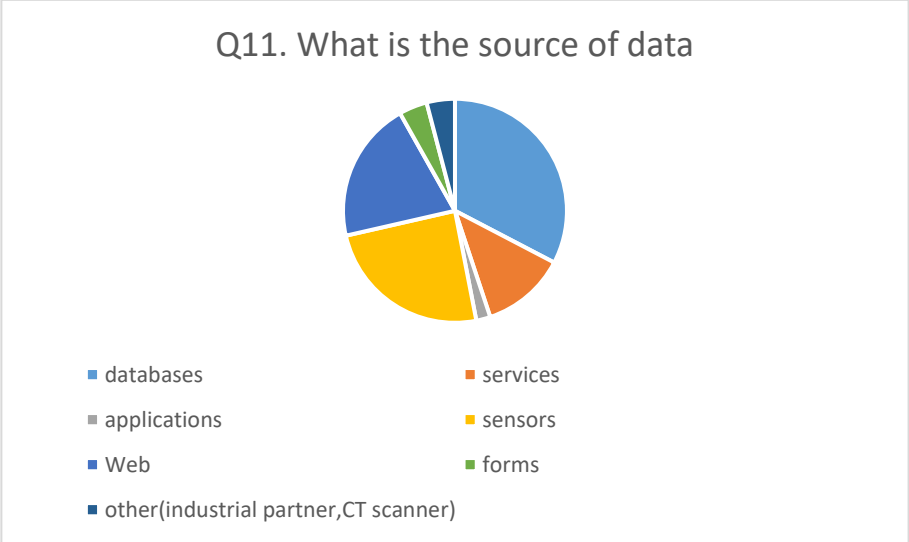
projektach. Ten rodzaj licencjonowania jest powszechny w branży oprogramowania, ale często ogranicza sposoby, w jakie użytkownicy mogą używać i rozpowszechniać oprogramowanie. Niektóre firmy nadal decydują się na korzystanie z zastrzeżonych modeli licencjonowania jako sposobu na ochronę swojej własności intelektualnej i utrzymanie kontroli nad swoim oprogramowaniem. Licencje copyleft nie są tak powszechne jak licencje permisywne lub własnościowe, ponieważ nakładają więcej ograniczeń na sposób używania i rozpowszechniania oprogramowania

Dyskusja:

- Licencje permisywne są popularne, ponieważ oferują równowagę między swobodą a elastycznością, która pozwala na współpracę i innowacje, jednocześnie minimalizując bariery wejścia i adopcji.
- Licencje copyleft są mniej powszechne niż licencje permisywne lub własnościowe, ponieważ nakładają więcej ograniczeń, co może nie być pożądane dla wszystkich użytkowników i programistów.

3.9 Jakie jest źródło danych?

Pytanie 11 zawiera odpowiedzi dotyczące źródła danych. Opcje różnią się od baz danych, usług, aplikacji, czujników, sieci Web, formularzy lub innych (jeśli odpowiedź jest inna, respondent wprowadza źródło). Poniższy wykres kołowy przedstawia odpowiedzi i ich rozkład.



Opis danych:
Wyniki pokazują, że podstawowym źródłem danych są bazy danych (w 16 raportach), następnie czujniki (w 12 raportach) i Web (w 10 raportach). Najrzadziej spotykanym źródłem są formularze (w 2 raportach), inne, takie jak partner przemysłowy czy tomograf komputerowy (w 2 raportach) oraz aplikacje (w 1 raporcie). Bazy danych są preferowanymi źródłami, ponieważ dane w nich zawarte są ustrukturyzowane i można je efektywnie pobierać. Ponadto dane są spójne i dokładne, a bazy danych

można łatwo integrować z innymi systemami, co ułatwia wymianę informacji między różnymi aplikacjami i platformami. Z drugiej strony aplikacje i formularze nie są powszechnym źródłem informacji, ponieważ dane są często rozproszone, a ich jakość niespójna. Aplikacje lub formularze mogą nie być przeznaczone do obsługi dużych ilości danych, a dostęp do nich może być ograniczony ze względów bezpieczeństwa lub prywatności. Ponadto formularze mogą korzystać z różnych formatów danych i protokołów, co może utrudniać integrację danych między różnymi systemami i aplikacjami.

Dyskusja:

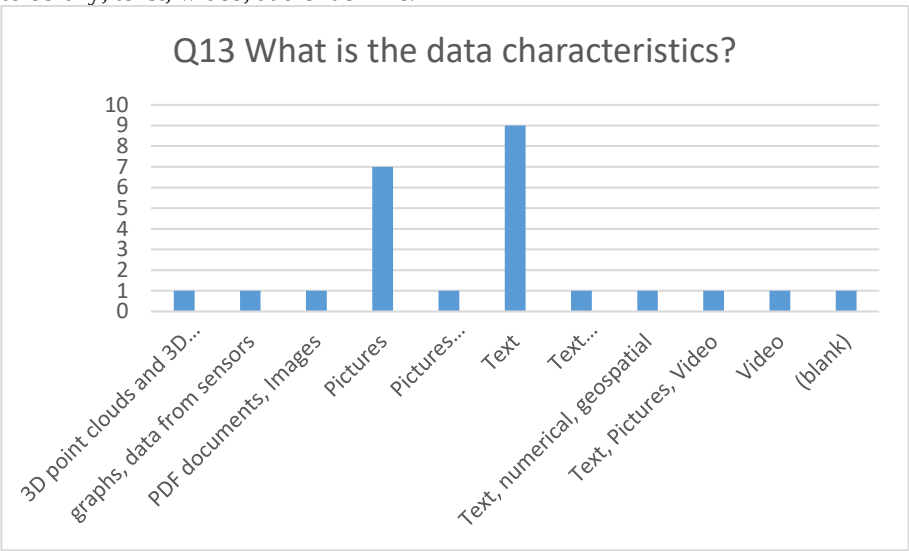
- Bazy danych są preferowanym źródłem danych, ponieważ oferują niezawodny, wydajny i bezpieczny sposób przechowywania dużych ilości danych i zarządzania nimi w spójny i ustrukturyzowany sposób.
- Aplikacje i formularze mogą być źródłami danych, ale często wymagają znacznego wysiłku i zasobów, aby je skutecznie wyodrębnić i wykorzystać.

3.10. Reprezentacja danych.

To pytanie zawiera informacje o typie danych. W projektach uwzględniono następujące typy - wykresy, dane z czujników, tekst, obrazy, csv, avro, parkiet, JSON API, serwer MySQL, serwer FTP, chmura punktów 3D, tomografia komputerowa klatki piersiowej, badania medyczne**.**

3.11. Jaka jest charakterystyka danych?

Poniższe pytanie zawiera informacje o charakterystyce danych. Możliwe odpowiedzi to obrazy, tekst, wideo, audio lub inne.



Opis danych:

Jak widać na wykresie, głównymi cechami danych są teksty i obrazy lub obrazy z tekstem lub wideo. Obrazy i teksty są łatwe do zrozumienia i interpretacji. Mogą być generowane w dużych ilościach i mogą być wykorzystywane w różnych kontekstach, od marketingu i reklamy po badania naukowe i analizę danych. Zdjęcia i teksty są łatwo dostępne i udostępniane na różnych platformach.

Dyskusja:

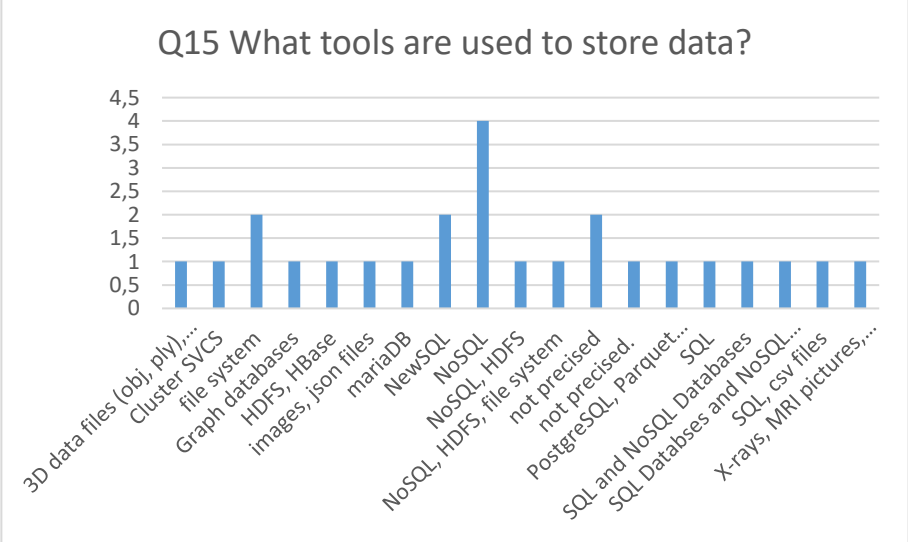
- Obrazy i teksty są typowymi cechami danych, ponieważ są łatwe do zrozumienia, generowane w dużych ilościach, wszechstronne i dostępne. Postęp technologiczny ułatwił również analizowanie i wydobywanie spostrzeżeń z nieustrukturyzowanych danych, co czyni je cennym źródłem informacji zarówno dla firm, jak i badaczy.

3.12. Przetwarzanie i jakość danych.

Pytanie 14 zawiera informacje na temat jakości danych i sposobu ich przetwarzania. W zależności od przypadku zastosowano różne metody. W niektórych projektach decyzje eksperckie były podejmowane w oparciu o wiedzę medyczną. W innych projektach dostęp do plików był uzyskiwany i formatowany za pomocą Pythona. W niektórych projektach dane zostały oczyszczone i zwizualizowane. W innych projektach zastosowano klasyfikację i regresję. W niektórych przypadkach zastosowano skalowanie, etykietowanie, przetwarzanie audio lub wideo. Techniki te okazały się najbardziej powszechne w przetwarzaniu danych**.**

3.13. Jakie narzędzia są używane do przechowywania danych?

To pytanie jest otwarte i zawiera informacje o narzędziach, które są używane do przechowywania danych w różnych projektach. Poniższy wykres przedstawia odpowiedzi.



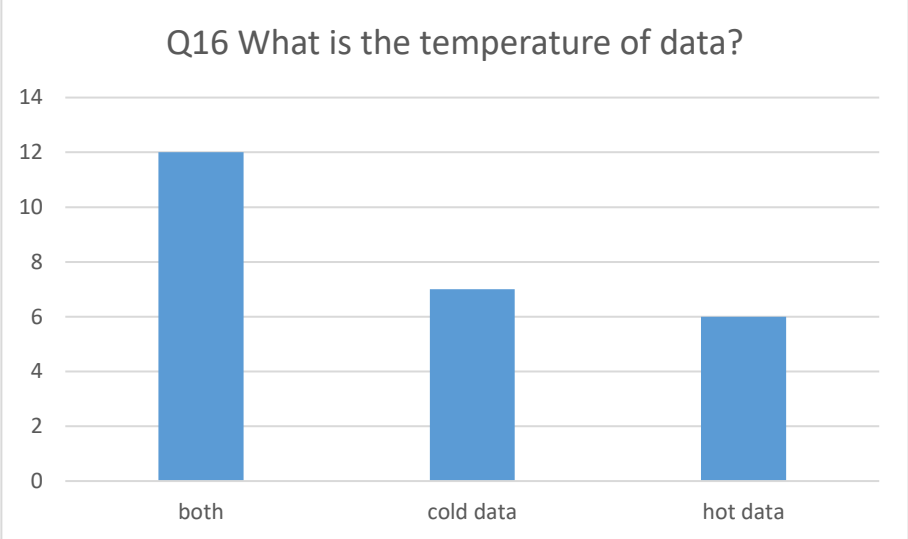
Opis danych:
Najpopularniejszymi narzędziami są bazy danych NoSQL. Te bazy danych są przeznaczone do obsługi danych nieustrukturyzowanych lub częściowo ustrukturyzowanych, co czyni je lepszym wyborem niż tradycyjne relacyjne bazy danych dla aplikacji wymagających elastycznego modelowania danych. Dzięki temu programiści mogą przechowywać i pobierać dane w sposób, który lepiej odpowiada potrzebom ich aplikacji. Wiele baz danych NoSQL jest typu open source, co czyni je opłacalną opcją dla programistów i organizacji. Pozostałe narzędzia są rozmieszczone niemal równo, więc można stwierdzić, że narzędzie, które zostanie użyte, zależy od przypadku i problemu informatycznego.

Dyskusja:

- Ogólnie rzecz biorąc, bazy danych NoSQL mają kilka zalet w porównaniu z tradycyjnymi relacyjnymi bazami danych i dlatego stają się coraz bardziej powszechne jako narzędzia do przechowywania danych. Oferują większą elastyczność, skalowalność, dostępność i wydajność, co czyni je idealnym wyborem dla nowoczesnych aplikacji, które wymagają tych funkcji.

3.14. Jaka jest temperatura danych?

Poniższe pytania pokazują, czy zimne, gorące, czy oba te dane są bardziej powszechne. Poniższy wykres przedstawia rozkład tych typów.



Opis danych:
Można wyraźnie zaobserwować, że zarówno zimne, jak i gorące dane. Przechowywanie zimnych danych jest zazwyczaj tańsze, ponieważ często są przechowywane na wolniejszych i tańszych urządzeniach pamięci masowej, takich jak taśmy lub dyski twarde. To sprawia, że jest to idealne rozwiązanie do przechowywania danych historycznych, kopii zapasowych i archiwów, do których

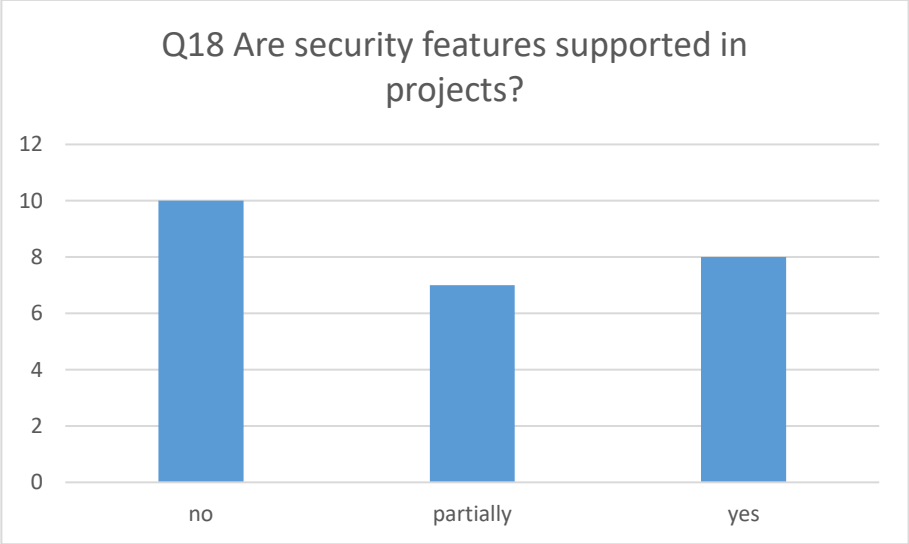
może nie być potrzebny częsty dostęp, ale muszą być przechowywane ze względu na zgodność lub ze względów prawnych. Z drugiej strony gorące dane są zwykle przechowywane na szybszych i droższych urządzeniach pamięci masowej. Tego typu dane są często dostępne dla aplikacji, użytkowników lub usług i muszą być szybko dostępne, aby obsługiwać operacje, transakcje lub analizy w czasie rzeczywistym. W większości organizacji zarówno gorące, jak i zimne dane są niezbędne do prowadzenia działalności biznesowej.

Dyskusja:

- Zarówno zimne, jak i gorące dane są powszechne w równym stopniu, ponieważ służą różnym celom i są potrzebne w różnym czasie. Zimne dane odnoszą się do danych, do których dostęp jest rzadki lub wcale, podczas gdy gorące dane odnoszą się do danych, do których często uzyskuje się dostęp lub które są aktywnie używane.

3.15. Czy w projektach obsługiwane są zabezpieczenia?

Następne pytanie dotyczy tego, czy funkcje zabezpieczeń są obsługiwane w projektach. Alternatywy to tak, nie lub częściowo**.**



Opis danych:
Wyniki wskazują, że nieco częściej funkcje bezpieczeństwa nie są obsługiwane niż odwrotnie. Możliwym wyjaśnieniem może być to, że wdrożenie solidnych funkcji zabezpieczeń może wymagać czasu i zasobów oraz może wymagać znacznych inwestycji w sprzęt, oprogramowanie i personel. W niektórych przypadkach organizacje mogą przedkładać inne funkcje nad bezpieczeństwo ze względu na ograniczenia budżetowe lub czasowe. Innym prawdopodobnym powodem może być brak wystarczającej wiedzy specjalistycznej, ponieważ bezpieczeństwo może być złożoną i wyspecjalizowaną dziedziną, a nie wszyscy programiści lub organizacje mogą mieć zasoby lub wiedzę, aby właściwie rozwiązać problemy związane z

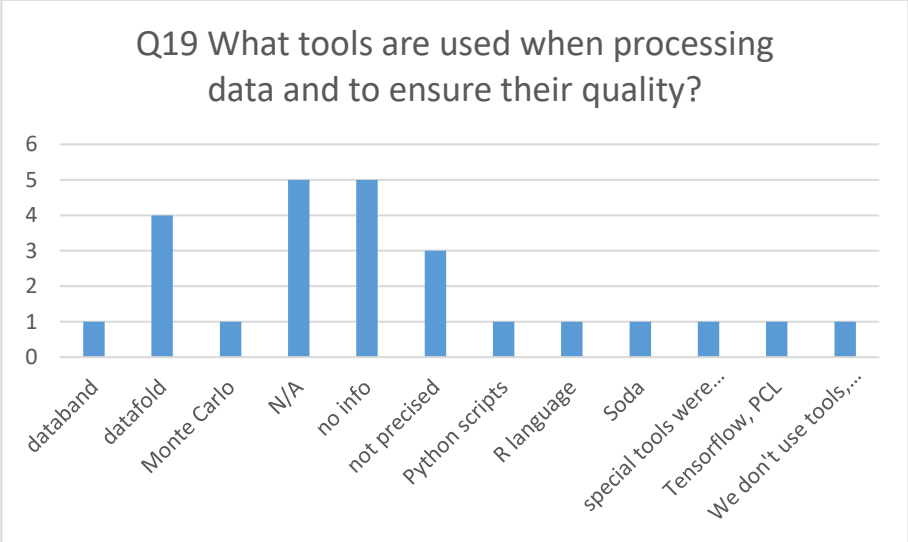
bezpieczeństwem. Jednak różnice są niewielkie, więc nie można uogólniać, że funkcje zabezpieczeń nie są obsługiwane, tylko w niektórych przypadkach.

Dyskusja:

- Funkcje zabezpieczeń nie są obsługiwane w większości projektów.
- Ważne jest, aby organizacje i deweloperzy nadali priorytet bezpieczeństwu i podjęli kroki w celu wdrożenia solidnych środków bezpieczeństwa w celu ochrony użytkowników i ich danych.

3.16. Jakie narzędzia są wykorzystywane przy przetwarzaniu danych i w celu zapewnienia ich jakości?

To pytanie sugeruje różne opcje, jeśli chodzi o narzędzia przetwarzania danych. Dostępne opcje to Talend, Toro, Soda, Datafold, Databand, Precisely, Monte Carlo lub inne. Rozkład wyników przedstawia poniższy wykres.



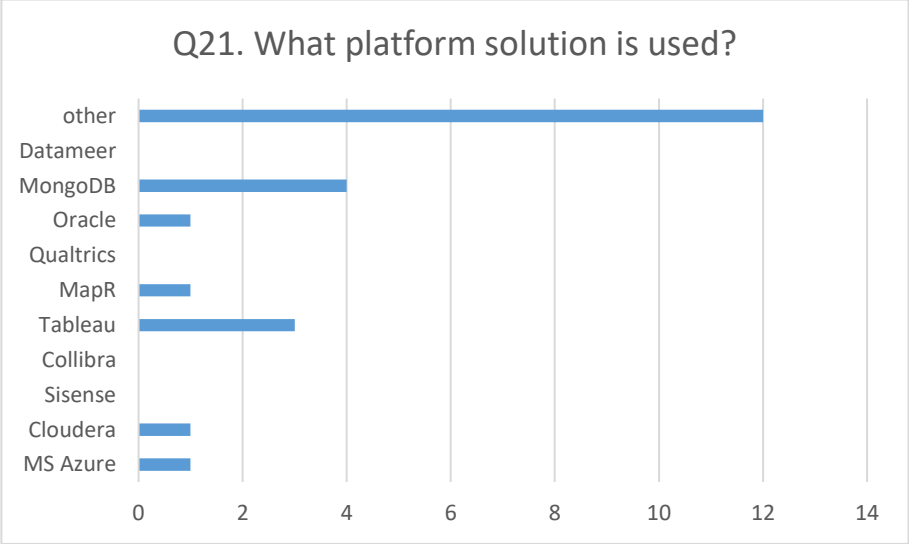
Opis danych:
Wyniki są w przeważającej mierze "nie dotyczy" lub "brak informacji", co sugeruje, że w analizowanych badaniach naukowcy nie wspomnieli o narzędziach, których używali podczas przetwarzania danych. Na trzecim miejscu znajduje się Datafold. Staje się coraz szerzej stosowany i jest uważany za cenne narzędzie do procesów przetwarzania danych.

Dyskusja:

- Narzędzia, które są wykorzystywane przy przetwarzaniu danych, nie są często wymieniane.
- Datafold to stosunkowo nowe narzędzie do przetwarzania danych, które zyskało popularność wśród inżynierów danych i naukowców zajmujących się danymi.

3.17. Jakie rozwiązanie platformowe jest stosowane?

Kolejne pytanie ujawnia, jakie rozwiązanie platformowe jest używane. Możliwe odpowiedzi to MS Azure, Cloudera, Sisense, Collibra, Tableau, MapR, Qualtrics, Oracle, MongoDB, Datameer lub inne, które wpisuje respondent. Wyniki przedstawiono na wykresie.



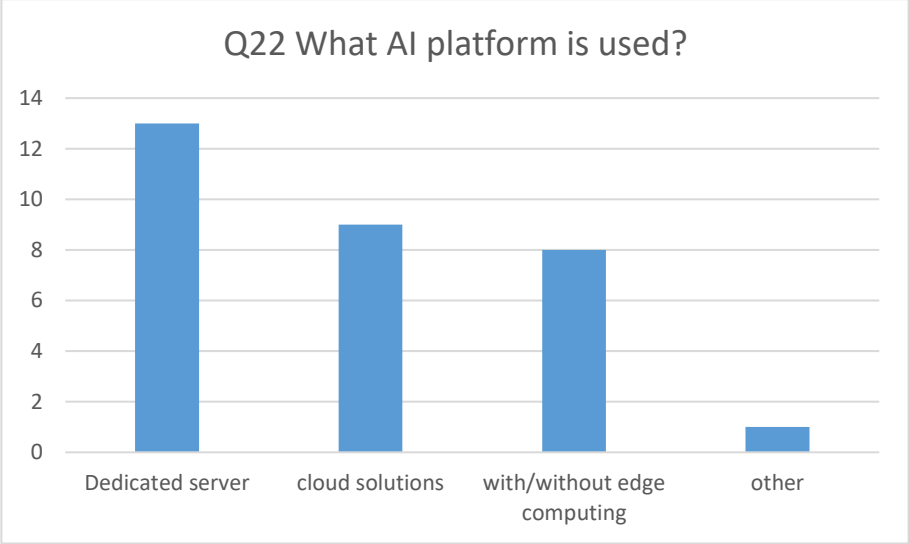
Opis danych:
Rozwiązania platformowe wykorzystywane w projektach to przede wszystkim inne. Należą do nich R studio, NVIDIA CUDA, Clara, CUDA-X, TensorFlow/TensorRT, Anaconda, Google Colab, Apache Spark, Apache Flink, PySpark, AWS, Bonseyes, specjalistyczny system informacji radiologicznej, Amazon Web Services S3 Data Lake, FedML, OpenMMLab, gRPC.

Dyskusja:

- Rozwiązania platformowe, które są sugerowane jako odpowiedzi, nie są tak popularne. MongoDB i Tableau zajmują odpowiednio drugą i trzecią pozycję.
- Inne platformy, takie jak Anaconda, Apache, NVIDIA, R studio, TensorFlow itp., są używane w większej liczbie przypadków.

3.18. Jaki rodzaj platformy sztucznej inteligencji jest wykorzystywany (np. rozwiązania oparte na serwerze, oparte na chmurze, z obsługą przetwarzania brzegowego lub bez niej lub inne)?

Poniższe pytanie dotyczy rodzaju platformy AI. Możliwe alternatywy to serwer dedykowany, rozwiązania chmurowe, z/bez obsługi przetwarzania brzegowego lub inne typy, wprowadzone przez respondentów. Wyniki przedstawiono na wykresie.



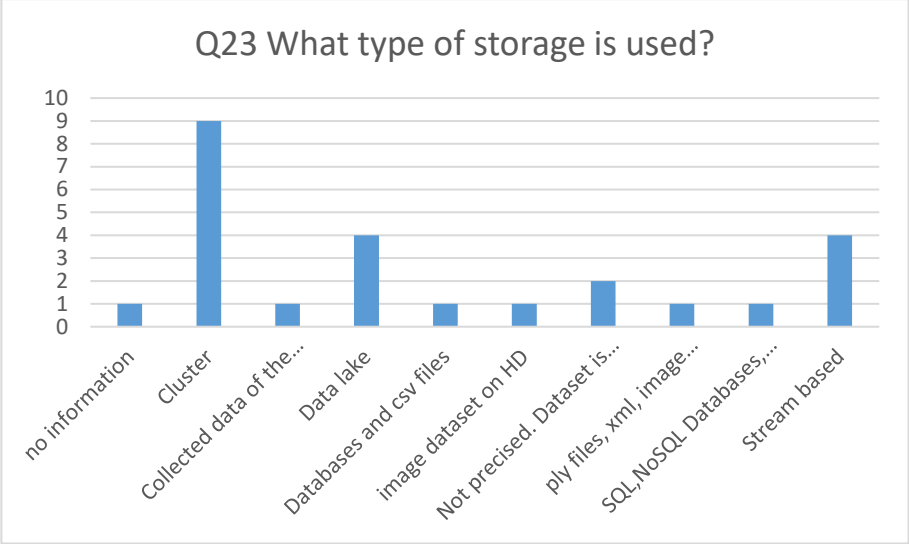
Opis danych:
Wyniki wskazują, że najczęściej używanym typem platformy AI jest serwer dedykowany. Serwery dedykowane są często używane jako typy platform AI, ponieważ oferują szereg zalet, które sprawiają, że dobrze nadają się do uruchamiania obciążeń AI. Serwery te można dostosować, ich wydajność jest wysoka i często są uważane za bezpieczniejsze niż współdzielone rozwiązania hostingowe. Ponadto serwery dedykowane można skalować w górę lub w dół i oferują większą kontrolę nad środowiskiem serwerowym niż współdzielone rozwiązania hostingowe.

Dyskusja:

- Wysoka wydajność, możliwość dostosowania, bezpieczeństwo, skalowalność i kontrola oferowane przez serwery dedykowane sprawiają, że są one popularnym wyborem do uruchamiania obciążeń AI.

3.19. Jaki rodzaj przechowywania jest używany?

Następne pytania zawierają szczegółowe informacje na temat typu magazynu. Sugerowane opcje to klaster, oparty na strumieniu, data lake lub inne. Rozkład odpowiedzi przedstawiono na wykresie.



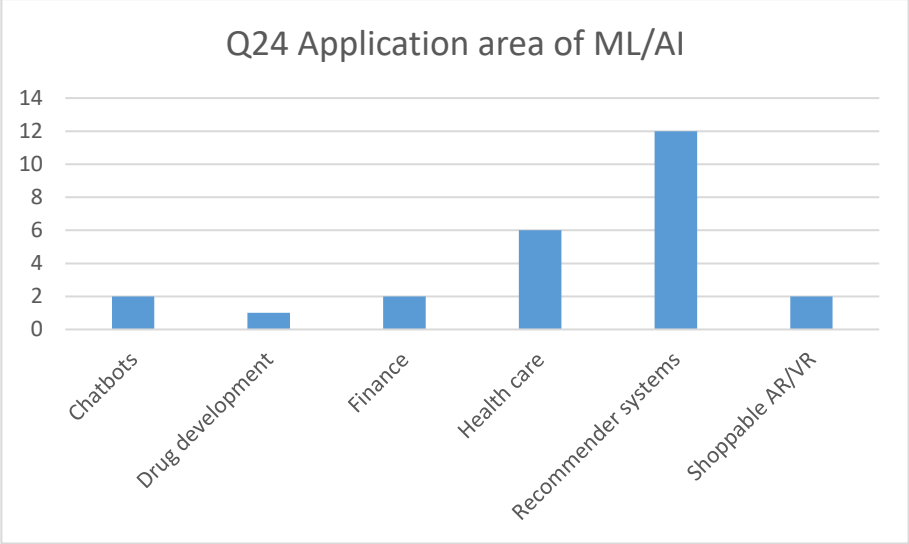
Opis danych:
Wyniki wskazują, że klaster jest najpopularniejszym rodzajem przechowywania wśród badaczy. Możliwym wyjaśnieniem może być to, że klastry są zaprojektowane w celu zapewnienia wysokiej dostępności danych. Klastry są wysoce skalowalne i odporne na uszkodzenia. Klastry mogą zapewniać wysoki poziom wydajności i mogą być ekonomicznym rozwiązaniem pamięci masowej, ponieważ można je tworzyć przy użyciu standardowego sprzętu.

Dyskusja:

- Klastry są popularne jako typ magazynu, ponieważ oferują wysoką dostępność, skalowalność, odporność na uszkodzenia, wydajność i opłacalność. Te zalety sprawiają, że są one idealnym wyborem dla firm i aplikacji, które wymagają niezawodnych i skalowalnych rozwiązań pamięci masowej.

3.20. Obszar zastosowania ML/AI?

Poniższe pytania ujawniają obszary zastosowań uczenia maszynowego/sztucznej inteligencji. Pytanie sugeruje osiem opcji – systemy rekomendacji, chatboty, testy A/B, rozszerzoną rzeczywistość z możliwością zakupu-AR/VR, opieka zdrowotna, rozwój leków, finanse, cyberbezpieczeństwo. Wyniki przedstawia poniższy wykres.



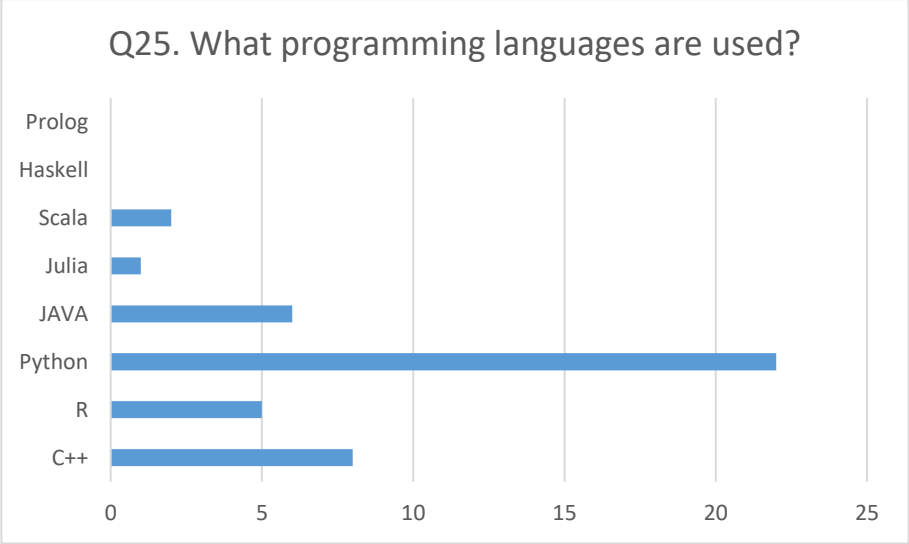
Opis danych:
Dane z analizowanych projektów pokazują, że najczęstszym obszarem, w którym ML i AI znajdują zastosowanie, są systemy rekomendacyjne. Systemy rekomendacji są zaprojektowane tak, aby dostarczać użytkownikom spersonalizowane rekomendacje na podstawie ich preferencji. Systemy rekomendacji zazwyczaj pracują z Big Data, takimi jak zachowania użytkowników i informacje o produktach. Wiele systemów rekomendacji jest wymaganych do dostarczania rekomendacji w czasie rzeczywistym, na przykład w przypadku witryn handlu elektronicznego lub platform streamingowych. Systemy rekomendacji mogą zapewnić znaczące korzyści biznesowe, takie jak zwiększona sprzedaż, zaangażowanie klientów i lojalność.

Dyskusja:

- Systemy rekomendacji są powszechnym obszarem zastosowania sztucznej inteligencji, ponieważ wymagają analizy złożonych zbiorów danych w celu dostarczania spersonalizowanych rekomendacji w czasie rzeczywistym. Algorytmy sztucznej inteligencji mogą pomóc zautomatyzować i usprawnić ten proces, co może prowadzić do lepszych wyników biznesowych i lepszych doświadczeń użytkowników.

3.21. Jakie języki programowania są używane?

To pytanie wskazuje języki programowania, które są używane w problemach AI. Wyniki przedstawiono na wykresie.



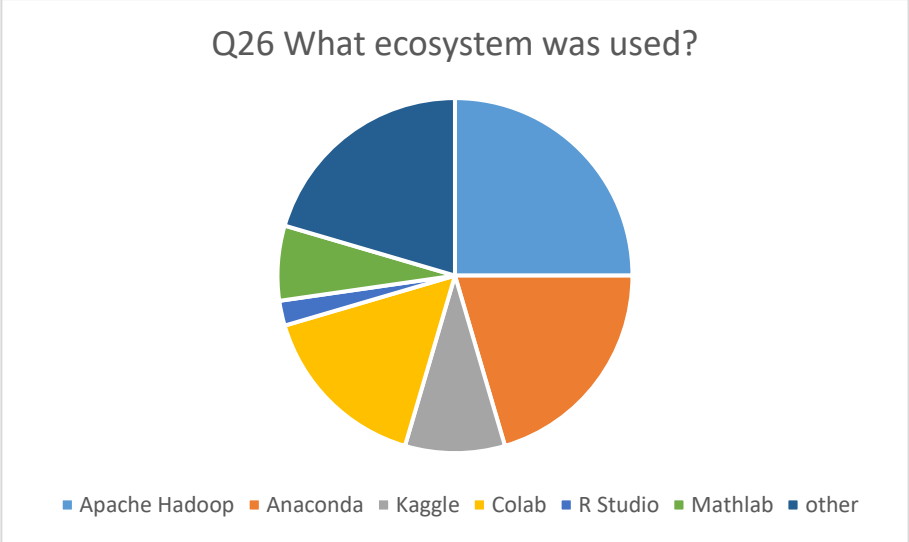
Opis danych:
Wyniki wyraźnie pokazują, że najczęściej używanym językiem programowania jest Python. Python ma dużą i aktywną społeczność programistów, którzy stworzyli liczne biblioteki i frameworki do rozwoju sztucznej inteligencji. Python to elastyczny język, który może być używany do szerokiego zakresu zadań związanych ze sztuczną inteligencją, w tym przetwarzania danych, uczenia maszynowego i przetwarzania języka naturalnego. Ponadto Python jest kompatybilny z szeroką gamą platform i systemów, w tym Windows, Mac i Linux. Chociaż Python nie jest najszybszym językiem programowania, jest wystarczająco szybki do większości zadań AI.

Dyskusja:

- Python jest popularnym językiem programowania dla sztucznej inteligencji ze względu na łatwość użycia, dużą społeczność, elastyczność, kompatybilność i wydajność. Te czynniki sprawiają, że jest to idealny wybór do tworzenia i wdrażania aplikacji AI.

3.22. Jaki ekosystem został wykorzystany?

Kolejne pytanie zawiera informacje o ekosystemie, który został wykorzystany. Oto możliwe alternatywy - Apache Hadoop, Anaconda, Kaggle, Colab, R studio, Mathlab lub inne. Rozkład odpowiedzi przedstawia poniższy wykres.



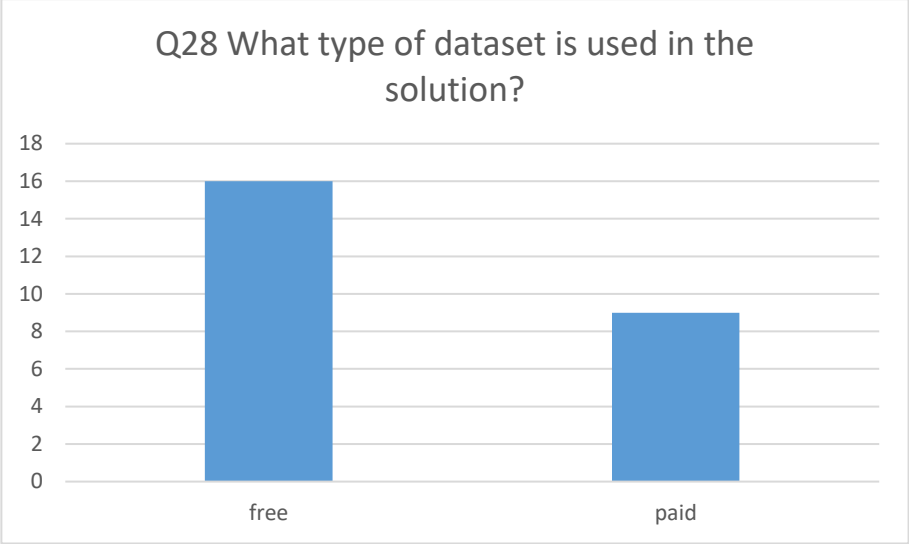
Opis danych:
Wyniki wskazują, że najczęściej używano Apache Hadoop, z niewielką przewagą nad Anacondą i innymi systemami. Opcja "inne"obejmuje ekosystemy, takie jak NVIDIA CUDA, TensorFlow, lokalne Python IDE, Apache Spark, DataBricks.

Dyskusja:

- Ekosystem Apache Hadoop jest szeroko stosowany do przetwarzania i analizy dużych zbiorów danych ze względu na jego skalowalność, przetwarzanie rozproszone, odporność na uszkodzenia, charakter open source, duży ekosystem i przyjęcie w branży. Czynniki te sprawiają, że jest to idealne rozwiązanie dla organizacji, które muszą przetwarzać i analizować duże ilości danych.
- Anakonda i inne ekosystemy związane głównie z językiem Python to również inne popularne ekosystemy. Anaconda to dystrybucja języków programowania Python i R wraz z kolekcją bibliotek typu open source, narzędzi i struktur do nauki o danych i obliczeń naukowych.

3.23. Jaki typ zbioru danych jest używany w rozwiązaniu?

Kolejne pytanie zawiera informacje o typie zbioru danych – czy jest on darmowy, czy płatny. Wyniki przedstawiono na wykresie kołowym.



Opis danych:
Wyniki wskazują na przewagę wolnych zbiorów danych, które są wykorzystywane w rozwiązaniach. Bezpłatne zestawy danych są bardziej dostępne dla szerszego grona użytkowników, w tym studentów, badaczy i programistów. Wiele bezpłatnych zbiorów danych to otwarte dane, co oznacza, że każdy może z nich swobodnie korzystać, modyfikować i rozpowszechniać. Płatne zestawy danych często wiążą się z ograniczeniami dotyczącymi sposobu ich wykorzystania, co może ograniczyć ich przydatność w przypadku niektórych typów analiz lub aplikacji.

Dyskusja:

- Dostępność bezpłatnych zbiorów danych przyczyniła się do wzrostu i rozwoju dziedzin nauki o danych i uczenia maszynowego poprzez promowanie dostępności, otwartości, współpracy i innowacji.

WNIOSKI

Tak więc kompleksowa analiza zebranych danych dostarcza cennych informacji na temat obecnego krajobrazu projektów sztucznej inteligencji (AI) i uczenia maszynowego (ML) podejmowanych przez różne instytucje badawcze. Najważniejsze wnioski można podsumować w następujący sposób.
Zebrano 25 kwestionariuszy z pięciu instytucji partnerskich. Większość szkoleń w zakresie sztucznej inteligencji odbywa się w Serbii, a następnie w Bułgarii. Badanie podkreśla potrzebę zwiększenia możliwości szkoleniowych, zwłaszcza w krajach Unii Europejskiej, w celu promowania dobrych praktyk w zakresie sztucznej inteligencji.

Deep Machine Learning (ML) jest najczęściej używanym modelem, demonstrującym swoją dominację w różnych zastosowaniach AI. Inne modele uczenia maszynowego, takie jak SciMl, są mniej rozpowszechnione wśród badanych projektów.
Konwolucyjne sieci neuronowe (CNN) są najczęściej opracowywanymi modelami, odzwierciedlającymi ich wszechstronność w zadaniach takich jak klasyfikacja obrazów, wykrywanie obiektów i analiza obrazów medycznych. W badanych projektach nie rozwinięto bramkowanej jednostki rekurencyjnej (GRU).
Najczęściej przetwarzane są ilości danych o pojemności od 1 GB do 1 TB, co jest zgodne z możliwościami nowoczesnych komputerów. Bardzo duże zestawy danych (ponad 1 TB) lub bardzo małe zestawy danych (mniej niż 1 GB) są mniej rozpowszechnione, co wskazuje na względy praktyczne i ograniczenia zasobów.
Rozwiązania AI są wdrażane w różnych dziedzinach, w tym w opiece zdrowotnej, finansach, inteligentnych miastach i Przemyśle 4.0. Każdy z respondentów zbadał konkretne obszary, zapewniając szeroką perspektywę zastosowań sztucznej inteligencji.
TensorFlow staje się najczęściej używaną biblioteką AI, a następnie Keras i Scikit-learn. Wybór biblioteki zależy od konkretnych potrzeb projektu, preferencji programistów i poziomu wiedzy.
Wyniki różnią się znacznie w zależności od projektu, obejmując takie zastosowania, jak wykrywanie ludzkiego ciała, analityka związana z COVID i wykrywanie obiektów. Każdy respondent odkrył unikalne wyniki dostosowane do konkretnego projektu.
Najbardziej rozpowszechnione są licencje permisywne, oferujące elastyczność i możliwości współpracy. Licencje własnościowe są drugim co do popularności, podczas gdy licencje copyleft nakładają więcej ograniczeń.
Bazy danych są podstawowym źródłem danych, zapewniającym uporządkowaną i wydajną pamięć masową.
Aplikacje i formularze są mniej powszechne ze względu na rozproszone dane i niespójną jakość.
Teksty i obrazy dominują jako cechy danych, odzwierciedlając ich łatwość zrozumienia, generowania i wszechstronności.
Do przetwarzania danych i zapewnienia jakości stosowane są różne metody, w tym decyzje eksperckie, formatowanie Pythona i klasyfikacja.
Bazy danych NoSQL są najczęściej używanymi narzędziami do przechowywania danych, oferującymi elastyczność i skalowalność.
Zarówno zimne, jak i gorące dane są powszechne, spełniając różne potrzeby i częstotliwości użytkowania.
Funkcje zabezpieczeń nie są spójnie obsługiwane w różnych projektach, a jako potencjalne przyczyny wymienia się ograniczenia zasobów i wiedzę specjalistyczną.
Wspomniane narzędzia do przetwarzania danych nie są dostarczane w sposób spójny, a jednym z nich jest Datafold.
Inne platformy, w tym R studio, NVIDIA CUDA i TensorFlow, dominują nad sugerowanymi opcjami, takimi jak MongoDB i Tableau.
Serwery dedykowane są najczęściej używanym typem platformy AI, oferującym wysoką wydajność, możliwości dostosowania i bezpieczeństwo.
Klastry są preferowanym typem magazynu, zapewniającym wysoką dostępność, skalowalność i odporność na uszkodzenia.

Systemy rekomendacji to najczęściej stosowane obszary ML/AI, wykorzystujące spersonalizowane rekomendacje.
Python jest zdecydowanie preferowanym językiem programowania dla projektów AI ze względu na wsparcie społeczności, elastyczność i kompatybilność.
Apache Hadoop jest najczęściej używanym ekosystemem, a następnie Anaconda i inne ekosystemy związane z Pythonem.
Dominują bezpłatne zestawy danych, które sprzyjają dostępności, otwartości, współpracy i innowacjom w dziedzinie sztucznej inteligencji i uczenia maszynowego.
Podsumowując, zróżnicowany zakres ustaleń podkreśla dynamiczny i ewoluujący charakter projektów związanych ze sztuczną inteligencją i uczeniem maszynowym, podkreślając znaczenie rozwiązań dostosowanych do potrzeb i ciągłego dostosowywania się do nowych technologii i metodologii. Zidentyfikowane trendy i wzorce stanowią podstawę dla przyszłych badań i rozwoju w szybko rozwijającej się dziedzinie sztucznej inteligencji.

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.