Raport științific
Etapa 2 – Proiectarea și implementarea componentelor

Universitatea POLITEHNICA din București, Facultatea de Automatică și Calculatoare,
Departamentul Calculatoare

CUPRINS

1 CONSIDERAȚII GENERALE 5

2 IDENTIFICAREA SCENARIILOR ȘI DEFINIREA CAZURILOR DE UTILIZARE 7

2.1 Water Monitoring System (WMS) 7

2.1.1 Scenariul NodeMCU (UTCN.C1) 9

2.1.2 Scenariul Raspberry Pi (UTCN.C2) 10

2.1.3 Scenariul Nvidia Jetson Nano (UTCN.C3) 11

2.2 Sistemul de suport decizional (DSS) 11

2.2.1 Analiza datelor 12

2.2.2 Suport decizional pentru gestionarea resurselor de apă (UPB.DSS) 15

2.3 Sistemul de Mobile Crowdsensing (MCS) 17

2.3.1 Monitorizarea infrastructurii de apă prin crowdsensing (UPB.CS) 17

2.4 Sistemul Blockchain (B) 19

2.4.1 Sistem de recompense pe bază de Blockchain (UPB.B) 19

3 PROIECTAREA ARHITECTURII SISTEMULUI 21

3.1 Arhitectura componentei de monitorizare a consumului (WMS) 22

3.1.1 Arhitectura de colectare și procesare a datelor la UTCN (UTCN.C) 22

3.1.2 Arhitectura de colectare a datelor la UPB (UPB.C) 23

3.2 Arhitectura componentei de suport decizional (DSS) 24

3.3 Arhitectura componentei de Crowdsensing (CS) 25

3.4 Arhitectura componentei Blockchain (B) 26

4 ANALIZA DATELOR 28

4.1 Seturi de date existente 28

4.1.1 Helix 28

4.1.2 GECCO 28

4.1.3 Austin Water 29

4.2 Scenarii de prelucrare a datelor 29

4.2.1 Detecția anomaliilor (Scenariul UTCN.A1) 29

4.2.2 Predicția consumului de apă (Scenariul UTCN.A2) 32
4.2.3 Identificarea tiparelor de consum săptămânal (Scenariul UPB.A1) 36
4.2.4 Clasificarea consumatorilor după nivelul deviației standard și al geo-localizării (Scenariul UPB.A2) 40
4.2.5 Clasificarea consumatorilor după categorii predefinite (Scenariul UPB.A3) 44

5 PROIECTAREA COMPONENTELOR ȘI INTERACȚIUNILOR 48

5.1 Monitorizare consum (WMS) 48
5.1.1 Componenta de monitorizare debit de apă (UTCN.C) 48
5.1.2 Proiectarea modulelor de achiziție (UPB.C) 49
5.1.3 Componenta Edge de preprocesare a datelor (UTCN.C) 49
5.1.4 Serviciul de colectare a datelor (UTCN.C) 50
5.1.5 Aplicația mobilă (UPB.WMS) 54

5.2 Suport decizional (DSS) 55
5.2.1 Serviciile de detecție a anomaliilor (UTCN.A1) și de predicție (UTCN.A2) 55
5.2.2 Clustering pe bază de date de consum folosind K-Means (UPB.A1) 57
5.2.3 Clustering pe bază de coordonate geografice folosind OPTICS (UPB.A2) 57
5.2.4 Metode de clasificare (UPB.A3) 58

5.3 Crowdsensing (MCS) 59
5.3.1 Model de alocare a raportărilor bazat pe VRP 59
5.3.2 Model de licitație bazat pe VCG pentru raportare veridică 61

5.4 Blockchain (B) 62
5.4.1 Truffle 62
5.4.2 Ganache 62
5.4.3 Smart Contracts 63
5.4.4 Metamask si Web3.js 63

6 IMPLEMENTAREA COMPONENTELOR SISTEMULUI 65

6.1 Monitorizare consum (WMS) 65
6.1.1 Componentele de monitorizare debit de apă și nivel edge (UTCN.C) 65
6.1.2 Serviciul de colectare a datelor (UTCN.C) 68
6.1.3 Configurarea modulelor de achiziție (UPB.C) 70
6.1.4 Interfață de comunicare (UPB.WSN) 71
6.1.5 Procesul de instalare (UPB) 73

6.1.6 Aplicație mobilă (UPB.WMS) ... 75
6.1.7 Serviciul de date (UPB.REST) ... 77
6.2 Suport decizional (DSS) ... 79
6.2.1 Serviciul de procesare/analiză a datelor (UTCN.A) ... 79
6.2.2 Serviciul de analiză a datelor (UPB.A) ... 85
6.3 Crowdsensing (CS) ... 88
6.4 Blockchain (B) ... 89
6.4.1 Aplicatie Web – Oferta de moneda inițială ... 89
6.4.2 Aplicație Web – Transfer criptomonedă ... 90
6.4.3 Funcționalitate REST ... 91
7 PREGĂTIREA DATELOR ... 92
7.1.1 Metode de procesare în edge (UTCN.C) ... 92
7.1.2 Metode de preprocesare premergătoare analizei datelor (UTCN.C) ... 92
7.1.3 Metode de stocare și vizualizare (UPB) ... 93
8 CONCLUZII ... 97
BIBLIOGRAFIE ... 98

1 Considerații generale

În contextul măsurilor de protecție a mediului și de combatere a fenomenelor asociate încălzirii globale, managementul infrastructurilor de furnizare a apei potabile joacă un rol primordial. Măsurile de gestionare eficientă a distribuției de apă și cele care vizează reducerea pierderilor și a consumurilor nerașionale, contribuie la menținerea unui echilibru stabil între producție și consum, iar la nivel global contribuie la menținerea unui echilibru în ecosistemul unei regiuni.

Din acest motiv sistemele informatice destinate pentru monitorizarea și optimizarea consumurilor de apă pot contribui semnificativ la menținerea acestui echilibru și asigurarea sustenabilității pe termen mediu și lung a unei anumite regiuni. Cantitatea și calitatea apei furnizate de o infrastructură de furnizare a apei potabile sau industriale influențează calitatea vieții pentru populația unei regiuni și asigură suportul pentru dezvoltarea unor activități economice productive din domeniul agricol și industrial.

În primul rând, colectarea datelor privind consumurile de apă și privind calitatea apei trebuie sa se facă la o rezoluție temporală și spațială care să permită identificarea anumitor profile de comportament pentru furnizorii și consumatorii de apă. Nu întâmplător în cadrul proiectului de față s-a urmărit dezvoltarea unor sisteme pilot care să identifice consumurile de apă din interiorul unei gospodării/apartament, lucru care, în infrastructurile actuale aflate în exploatare, nu se întâmplă. În locul colectării lunare a datelor de consum (în vederea facturării), în cadrul proiectului de față se promovează ideea unei monitorizări mult mai frecvente în mai multe puncte de consum. Acest lucru poate sa asigure suportul pentru identificarea mult mai rapidă a unor incidente în rețeaua de apă și, de asemenea, permite predicția cu o precizie mai mare a consumurilor de apă.

Dar, un management optimal al rețelei de furnizare a apei presupune, pe lângă partea de monitorizare și o parte de analiză și interpretare automată a datelor colectate în vederea detectării unor comportamente anormale. În acest scop, în cadrul proiectului s-au dezvoltat o serie de tehnici de analiză care vizează detecția de anomalii și predicția consumurilor viitoare. Pentru validarea și testarea metodelor propuse și implementate s-au folosit seturi de date oferite ca date de referință de către anumite proiecte anterioare. În mod concret s-au folosit două seturi de date:

- unul pentru analiza consumurilor de apă pe un set de apartamente/gospodării și pe o anumită perioadă de timp (Chatzigeorgakidis, 2018) și
- unul pentru detectarea de anomalii cu privire la calitatea apei (Chandrasekaran et al., 2017).

Rezultatele obținute în faza de analiză sunt destinate pentru asigurarea suportului decizional necesar în managementul în timp-real al rețelelor de apă.

Eficiența implementări unui sistem de management al unei rețele de apă depinde în egală măsură de factorii decizionali, dar și de contribuția și implicarea directă a consumatorilor. În acest scop, proiectul de față propune un sistem colaborativ de colectare a informațiilor bazat pe două concepte: Crowdsensing (colectare de date prin contribuția utilizatorilor) și Serious Gaming (introducerea de recompense pentru contribuțiile individuale, asemănător cu tehnicile utilizate în jocuri). În acest mod utilizatorii devin factori activi în optimizarea consumurilor de apă și în detectarea în timp util a incidentelor din rețea.

Raportul de față prezintă stadiul actual al cercetărilor și dezvoltările practice realizate în cadrul proiectului în etapa a doua a proiectului.

Astfel în următorul capitol (Capitolul 2) sunt identificați actorii sistemului și scenariile de utilizare. Capitolul 3 oferă o viziune generală asupra arhitecturii sistemului, evidențiindu-se componentele funcționale și intercondiționările existente între ele.

Capitolul 4 este destinat descrierii procedurilor utilizate în analiza de date. Astfel sunt descrise tehnicile de detecție a anomaliilor implementate, precum și metodele folosite în predicția consumurilor de apă, oferind detalii metodologice de procesare a datelor. Pentru partea de detecție a anomaliilor s-au implementat mai multe tehnici de inteligență artificială care permit detectarea automată a unor situații critice în rețeaua de apa. Platforma dezvoltată în acest scop permite identificarea automată a celui mai eficient algoritm de detecție a anomaliilor pentru un anumit set de date. Metricile de măsurare a calității rezultatelor obținute indică un nivel de calitate comparabil și uneori chiar superior rezultatelor raportate în literatura de specialitate.

Pentru predicția consumurilor de apă s-au folosit tehnici de procesare a seriilor de timp. Aceste metode exploatează corelațiile interne existente în datele colectate și, pe baza acestora, calculează valorile de consum pe o anumită fereastră de predicție.

Capitolul 5 cuprinde proiectarea detaliată a componentelor sistemului și a interacțiunilor dintre acestea. Pentru asigurarea scalabilității și extensibilității sistemului, aceste componente au fost concepute sub forma unor servicii accesibile prin interfețe standardizate.

Capitolul 6 oferă anumite detalii de implementare a componentelor prezentate în capitolul anterior, iar în Capitolul 7 sunt prezentate metodele de aplicare a algoritmilor de procesare a datelor prezentați anterior pe datele obținute în urma implementării componentelor sistemului.

Raportul se încheie printr-un capitol de concluzii, care sintetizează rezultatele obținute până în această etapă și trasează pașii de urmat în etapa următoare a proiectului.

2 Identificarea scenariilor și definirea cazurilor de utilizare

Proiectul are ca scop realizarea unui model de laborator pentru un sistem de tip CPSS (Cyber-Physical-Social System) pentru dezvoltare sustenabilă în domeniul furnizării de apă în mediul urban prin implicarea activă a comunității. În implementarea acestuia intervin mai multe scenarii determinate de factori ce țin de infrastructura inteligentă precum și de interacțiunea cu cetățeanul. Pentru a defini scenariile este important de avut în vedere contextul general stabilit în urma analizei stadiului actual și a posibilităților de dezvoltare.

Sistemul dezvoltat este de fapt un meta-sistem (Fig. 1), un ansamblul de componente ce interacționează între ele, fiecare componentă putând fi modelată printr-un set de cazuri de utilizare specifice. Arhitectura sistemului evidențiază componentele funcționale și intercondiționările dintre ele, iar detaliile de implementare a componentelor prezintă sistemele hardware și metodele software utilizate.

Am considerat că meta-sistemul dezvoltat trebuie să aibă 4 componente, reprezentate de sistemul de monitorizare a consumului (WMS), sistemul de suport decizional (DSS), sistemul colaborativ de raportare prin crowdsensing (MCS) și sistemul blockchain (B).

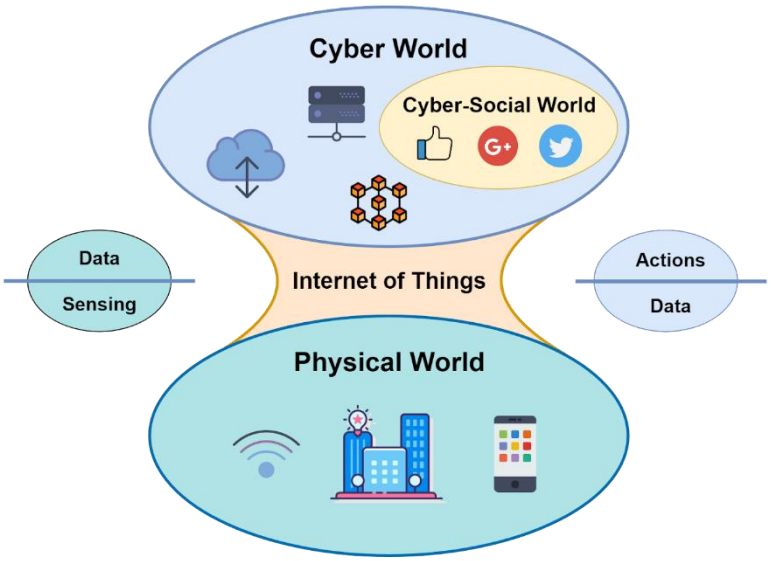


Fig. 1 Arhitectura generală de tip CPSS

2.1 Water Monitoring System (WMS)

Sistemul de monitorizare a consumului de apă poate fi descris prin următoarele scenarii generale:

- Instalarea și configurarea senzorilor la rețeaua de apă a cetățeanului;
- Colectarea și trimiterea datelor către server în timp real;
- Oferirea unei imagini de ansamblu asupra consumului de apă al cetățenilor, cu scopul de a realiza recomandări personalizate.

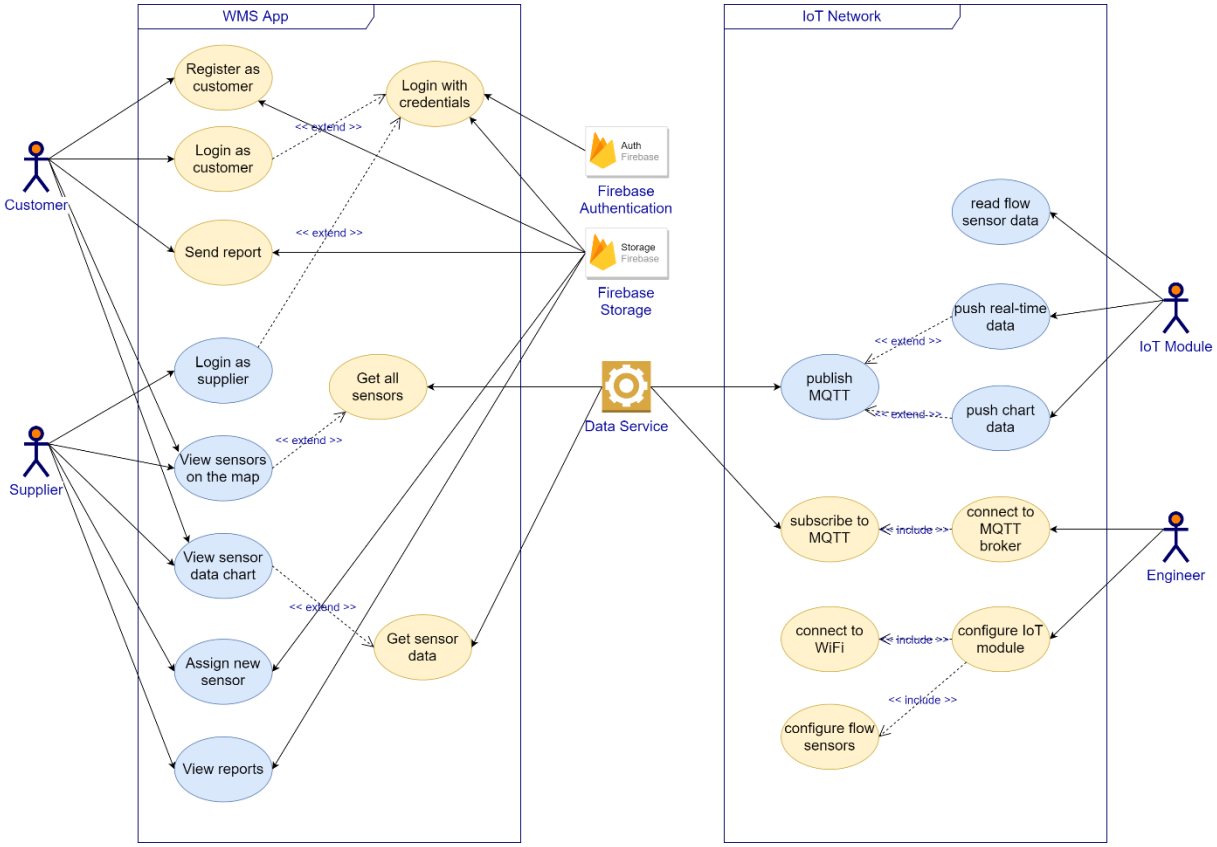


Fig. 2 Diagrama cazurilor de utilizare pentru sistemul de monitorizare a consumului de apă în locuințe (UPB.C)

Pentru definirea cazurilor de utilizare (a consumului de apă în locuințe) au fost considerate următoarele componente și interacțiuni (UPB.C):

- WMS App - aplicație monitorizare consum;
- IoT Network - rețea de senzori;
- Data Service - serviciu colectare a datelor de consum;
- Firebase Authentication - autentificare utilizatori;
- Firebase Storage – serviciu de stocare a datelor de la utilizatori.

Există mai multe tipuri de utilizatori considerați în sistemul de monitorizare a consumului de apă: consumator (Customer) și furnizor (Supplier), precum și modulele de achiziție a datelor de consum (IoT Module) și tehnicianul (Engineer) (Fig. 2). Consumatorul și furnizorul interacționează cu sistemul de monitorizare prin intermediul aplicației.

Astfel, aceștia se pot autentifica având roluri diferite și funcționalități specifice. Consumatorul poate să vizualizeze datele de la senzorii asociați, în timp ce furnizorul poate vizualiza datele de la toți senzorii instalați în locuințe. În plus, furnizorul poate realiza asocierea dintre senzori / module de achiziție și consumatori.

Funcțiile de autentificare și raportare (trimitere și vizualizare rapoarte) sunt procesate de serviciul Firebase. Datele de la senzori sunt colectate prin intermediul modulelor IoT care se conectează prin MQTT (Message Queuing Telemetry Transport) la serviciul de date.

Serviciul de date permite monitorizarea consumului de apă în timp real precum și colectarea acestora de la modulele de achiziție. Configurarea modulelor se realizează de către tehnician și include stabilirea conexiunii la rețeaua locală de WiFi și configurarea debitmetrelor atașate.

Pentru colectarea datelor și pre-procesarea acestora în edge se poate apela la 3 scenarii de utilizare diferite:

- NodeMCU (UTCN.C1) – datele sunt preluate de debitmetre și transmise cu ajutorul unui singur nod de tip NodeMCU în cloud;
- Raspberry Pi (UTCN.C2) - datele sunt preluate de debitmetre și transmise cu ajutorul unui / mai multor nod(uri) de tip NodeMCU către un dispozitiv de tip Raspberry Pi, care le pre-procesează și apoi le trimite în cloud;
- Nvidia Jetson Nano (UTCN.C3) - datele sunt preluate de debitmetre și transmise cu ajutorul unui / mai multor nod(uri) de tip NodeMCU către un dispozitiv de tip Nvidia Jetson Nano, care le pre-procesează și apoi le trimite în cloud;

2.1.1 Scenariul NodeMCU (UTCN.C1)

Datele cu privire la consum sunt colectate cu ajutorul debitmetrului care măsoară în litri / oră și care este conectat la un nod de tip NodeMCU cu posibilitate de transmitere wireless. Datele primare furnizate de către debitmetru sunt folosite la nivelul nodului pentru a se calcula consumul de apă și apoi sunt trimise cu frecvență orară spre serviciul web de colectare a datelor (Fig. 3).

Dacă există mai multe puncte de măsurare instalate în aceiași locație, din fiecare punct de măsurare, după calcularea datelor de consum, acestea se trimit individual către serviciul web de colectare a datelor (Fig. 4).

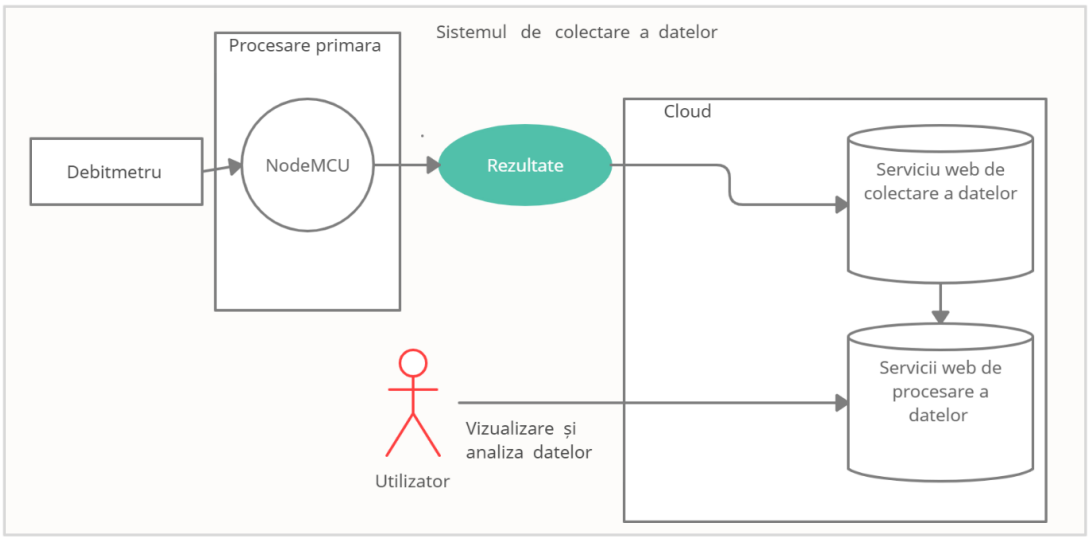


Fig. 3 Colectarea datelor neprocesate, direct de la punctele de monitorizare

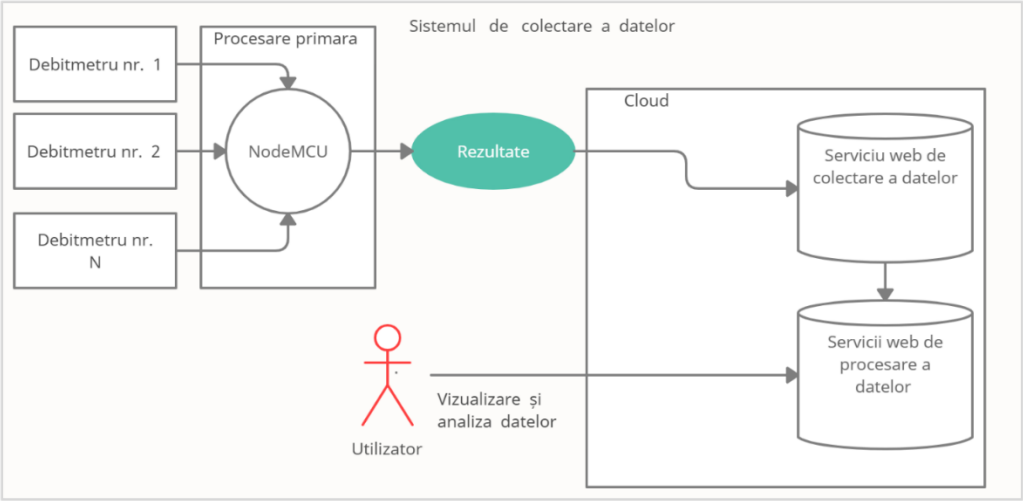


Fig. 4 Colectarea datelor preprocesate, direct de la punctele de monitorizare

2.1.2 Scenariul Raspberry Pi (UTCN.C2)

Datele cu privire la consum sunt colectate din mai multe puncte de măsurare din cadrul aceleiași locații cu ajutorul debitmetrelor care măsoară în litri / oră și care sunt conectate la unul sau mai multe noduri de tip NodeMCU cu posibilitate de transmitere wireless[1].

În fiecare locație este instalat un dispozitiv (de tip Raspberry Pi) care colectează datele transmise de către nodurile NodeMCU de la dispozitivele de măsurare a debitului (debitmetre), pe care apoi le procesează (calculând consumul de apă pe oră, elimină valorile extreme, calculând consumul total pe întreaga locație etc.), apoi transmite datele finale către serviciul de colectare a datelor (Fig. 5).

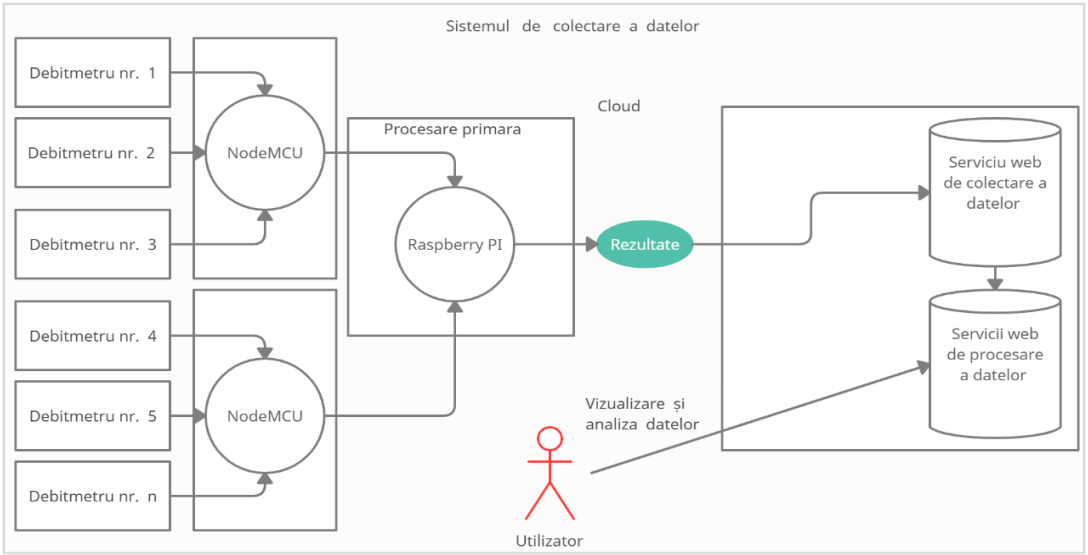


Fig. 5 Colectarea datelor preprocesate de la dispozitiv edge

[1] https://www.espressif.com/sites/default/files/documentation/0a-esp8266ex_datasheet_en.pdf

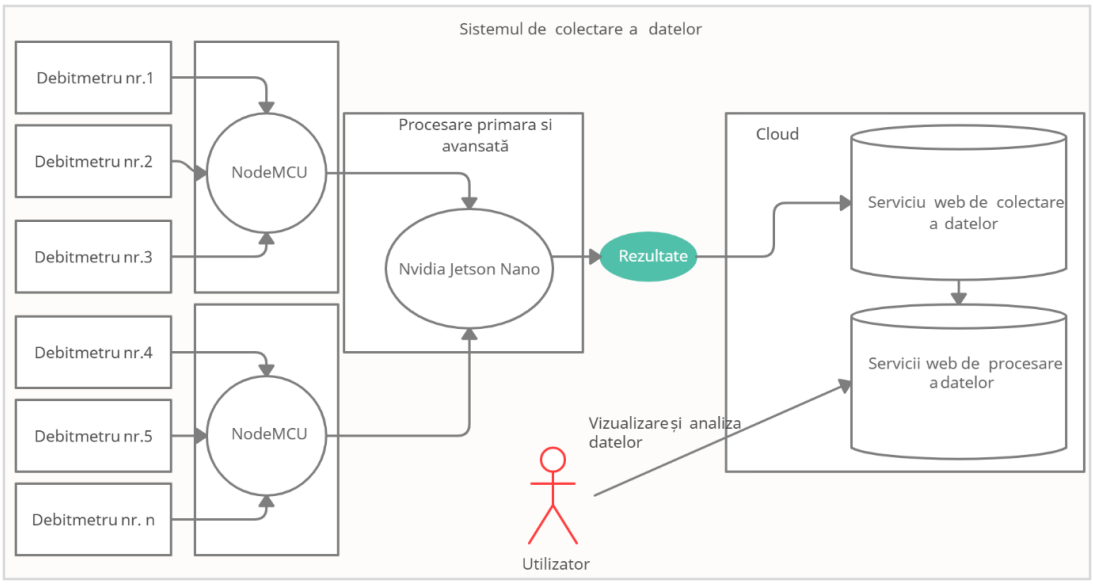


Fig. 6 Colectarea datelor preprocesate și etichetate de la dispozitiv edge

2.1.3 Scenariul Nvidia Jetson Nano (UTCN.C3)

Datele cu privire la consum sunt colectate din mai multe puncte de măsurare din cadrul aceleiași locații cu ajutorul debitmetrelor care măsoară în litri / oră și care sunt conectate la unul sau mai multe noduri de tip NodeMCU cu posibilitate de transmitere wireless.

În fiecare locație este instalat un dispozitiv cu capacitate avansată de procesare a datelor (model Nvidia Jetson Nano[2]) care colectează datele primare transmise de către nodurile NodeMCU la care sunt conectate dispozitivele locale de măsurare a debitului și le etichetează pe baza unui algoritm de învățare automată. Datele etichetate sunt trimise către serviciul web de colectare (Fig. 6).

2.2 Sistemul de suport decizional (DSS)

Sistemul de suport decizional presupune analiza datelor și apoi utilizarea arborilor și a sistemelor de decizie pentru a oferi recomandări în vederea optimizării consumului și a costurilor implicate, precum și educarea utilizatorilor. De asemenea, acest sistem este responsabil de realizarea de diverse vizualizări a consumatorilor pentru a ușura luarea de decizii.

Sistemul de suport decizional poate fi descris prin următoarele două scenarii generale: Analiza datelor și Suport decizional pentru gestionarea resurselor de apă.

2 https://developer.download.nvidia.com/assets/embedded/secure/jetson/Nano/docs/JetsonNano_DataSheet_DS09366001v1.0.pdf?RtjxplJRTx_QJRZxSUB3_Q2WmInp04kWBDDt_o4nNDgMOMnBSO80t0NBJ5JSRe835i0FkAnGKNM0gz1fn40YeR2Kor2ZJNgzFQ0MpJmxBVy_84JILc0WKCNzc891FQHv7nweD3VttK8DA50_S2E0HsEhcdW5vs1QdRdwKKmV_c1WqQW2GsyhS1CP0hEeyg&t=eyJscyI6IndlYnNpdGUiLCJsc2QiOiJkZXZlbG9wZXIubnZpZGlhLmNvbVwvcmVzdHJpY3RlZD9maWxlbmFtZT00MDMuaHRtbCJ9

2.2.1 Analiza datelor

Analiza datelor presupune prelucrarea datelor ce au fost preluate în mod automat de Data Service (serviciul de colectare a datelor de consum) sau urcate manual în cloud spre a fi prelucrate. Pentru analiza datelor s-au prevăzut cinci scenarii de utilizare:

- Detectarea anomaliilor (UTCN.A1);
- Predicția consumului (UTCN.A2);
- Identificarea tiparelor de consum săptămânal (UPB.A1);
- Clasificarea consumatorilor după nivelul deviației standard și al geo-localizării (UPB.A2);
- Clasificarea consumatorilor după tipare predefinite (UPB.A3).

Scenariul Detectarea anomaliilor (UTCN.A1)

În cadrul acestui scenariu se încearcă detecția anomaliilor în date de calitatea apei culese din rețeaua de distribuție, presupunându-se că acestea au fost colectate și stocate anterior, iar analiza lor ale loc off-line. Pentru experimente s-a folosit setul de date GECCO (Chandrasekaran et al., 2017).

Scenariul Predicția consumului (UTCN.A2)

Cel de-al doilea scenariu presupune predicția valorilor de consum a apei. În cadrul acestui scenariu se presupune că datele de consum a apei sunt colectate de la mai multe gospodării. În urma analizei datelor deja colectate, se prezic valorile de consum pentru zilele următoare. Setul de date folosit pentru experimente a fost Helix (Chatzigeorgakidis, 2018).

Cazurile de utilizare pentru aceste scenarii sunt enumerate mai jos și ilustrate în Fig. 7.

- Achiziționarea datelor:
  - Prin încărcarea fișierelor – fișiere .csv;
  - Prin conectare la diferite baze de date – MongoDB;
- Crearea pipeline-ului dinamic de procesare și analiza datelor încărcate:
  - Preprocesarea datelor:
    - Curățarea datelor – tratarea datelor lipsă;
    - Transformarea datelor – scalare și normalizare;
    - Reducerea datelor sau reechilibrarea setului de date – algoritmii SMOTE (Chawla et al., 2002) și RUID (Prusa et al., 2015).
  - Detecția anomaliilor în calitatea apei:
    - Selecția algoritmului de clasificare, câte opțiuni fiind:
      - kNN (Cunningham & Delany, 2007);
      - SVM (Evgeniou, & Pontil, 1999);
      - Decision Tree (Topîrceanu, 2017);
      - Random Forest (Breiman, 2001);
      - Extra Tree Classifier (Geurts, Ernst, & Wehenkel, 2006);
      - Logistic regression (Alzen, Langdon, & Otero, 2018);
      - Etc.
    - Antrenarea modelului.
    - Evaluare și vizualizare rezultate:

- Metrici de clasificare, cum ar fi acuratețe, precizie, scor f1;
- Generarea matricei de confuzie (confusion matrix);
- Generarea istoricului de antrenare.

  - Predicția valorilor de consum de apă.
- Vizualizarea seriilor de timp – line chart.
- Crearea modelului:
  - Selecția arhitecturii de LSTM:
    - LSTM (Hochreiter & Schmidhuber, 1997);
    - Encoder-Decoder LSTM (Lu et al., 2020);
    - CNN-LSTM Encoder-Decoder (Lu et al., 2020);
    - ConcLSTM-LSTM Encoder-Decoder (Xingjian et al., 2015).
  - Setarea parametrilor.
  - Antrenarea modelului.
  - Vizualizarea rezultatelor:
    - Vizualizarea valorilor prezise ca serii de timp;
    - Vizualizarea valorilor prezise în mod tabelar.

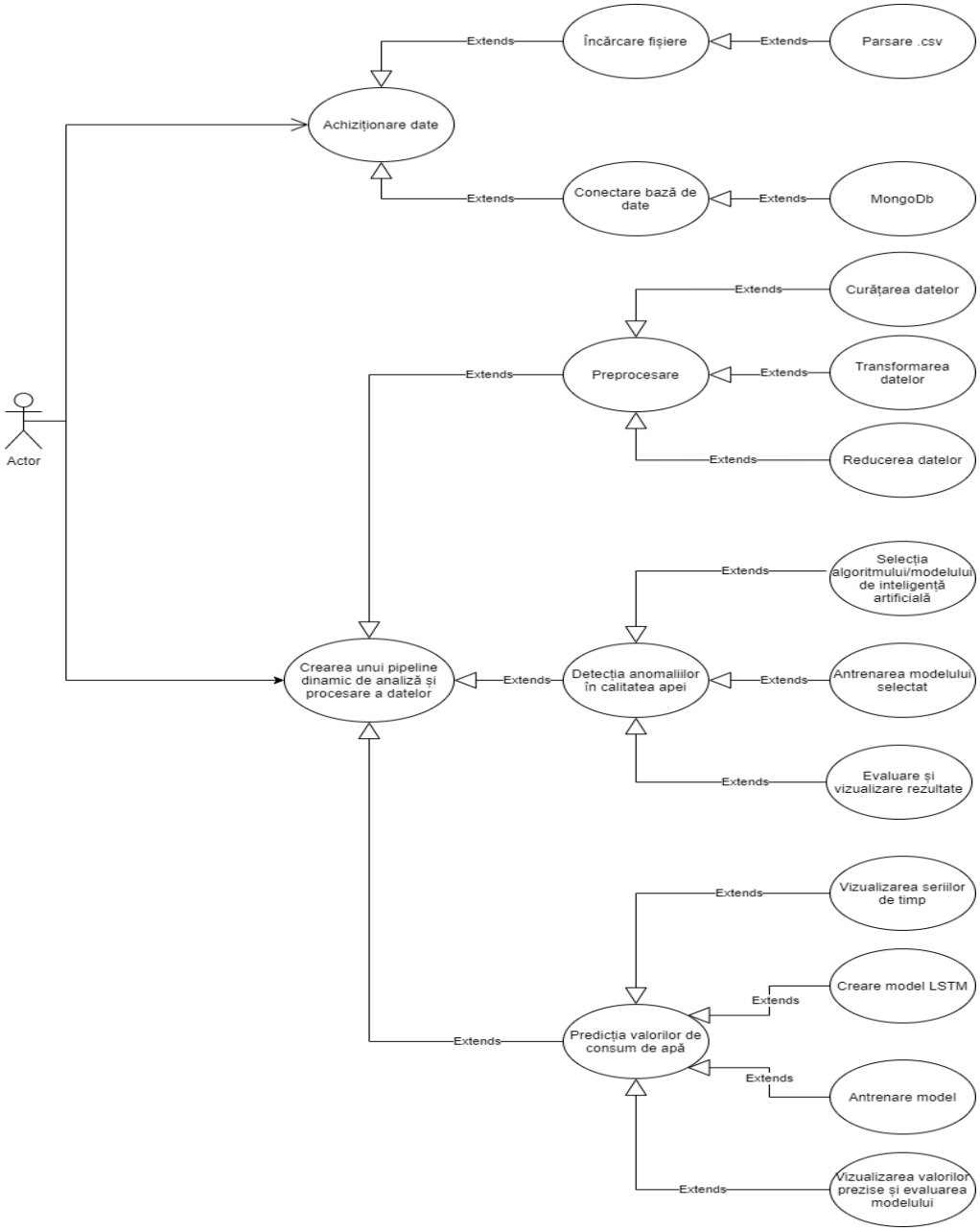


Fig. 7 Diagrama cazurilor de utilizare pentru procesare și analiză de date

Scenariul Identificarea tiparelor de consum săptămânal (UPB.A1)

Pentru cel de-al treilea scenariu, atenția s-a îndreptat asupra consumatorilor, în încercarea de a-i grupa în funcție de consumul de resurse de apă. Profilarea consumatorilor într-un sistem de distribuție a apei este esențială pentru realizarea sustenabilității în ceea ce privește gestionarea resurselor și dezvoltarea urbană. Învățarea nesupervizată poate oferi suport de decizie bazat pe date pentru evaluarea tiparelor cererii de apă într-o rețea de apa, în timp ce se pot adăuga diverse metode de pre-procesare pentru a extinde nivelul de detaliu în ceea ce privește comportamentul consumatorului. Experimentele corespunzătoare acestui scenariu au folosit tot setul de date Helix (Chatzigeorgakidis, 2018) prin aplicarea metodei de grupare (clustering) K-Means.

Continuând pe aceeași direcție, se propune identificarea cantității de variație pe perioada săptămânală pentru fiecare tipar de consum identificat. Pentru aceasta, s-a apelat mai întâi la normalizarea datelor, ceea permite evidențierea anumitor caracteristici precum:

- nivelul de variație a consumului pe unitatea de timp: reprezintă un factor ce influențează nivelul de uzură în timp a infrastructurii de apă, în funcție de care se vor calcula indicatorii de consum în vederea implementării sistemului de suport decizional pentru creșterea nivelului de sustenabilitate;
- nivelul de variație a consumului în raport cu ceilalți consumatori: exprimă o analiză calitativă, evidențiată prin forma tiparelor de consum, independentă de volumul efectiv consumat, fiind parte din sistemul de suport decizional în vederea generării unor recomandări în raport cu ceilalți consumatori din rețea.

Clasificarea consumatorilor după nivelul deviației standard și al geo-localizării (UPB.A2)

Cel de-al patrulea scenariu și-a propus clasificarea consumatorilor în funcție de zonele geografice unde locuiesc, identificate prin geo-locația acestora, precum și a nivelului de variație ce se poate observa în comportamentul lor. Descoperirea obiceiurilor consumatorilor este esențială pentru sprijinirea eficientă a deciziilor în rețelele inteligente de apă. În timp ce apometrele inteligente pot furniza date detaliate de consum pentru gospodăriile individuale, informații suplimentare pot fi extrase pe baza coordonatelor geografice, pentru a evidenția distribuția comportamentelor consumatorilor într-o anumită zonă. Nivelul deviației standard permite analiza cantitativă a nivelului de variație a consumului în timp ce geo-localizarea introduce un factor important în cazul grupurilor de consumatori.

În vederea optimizării consumului la nivel de rețea de distribuție, consumatorii sunt grupați în funcție de locație pentru a separa caracteristicile individuale de cele cumulate. Acestea din urmă reprezintă, de fapt, obiectivul principal în contextul sistemelor de distribuție la nivel de oraș. Prin calcularea nivelului deviației standard la nivel de consumatori și zone geografice, sistemul de suport decizional poate genera recomandări pentru îmbunătățirea eficienței în furnizarea apei, prin netezirea caracteristicilor de consum la nivel de rețea.

Clasificarea consumatorilor după tipare predefinite (UPB.A3)

În fine, ultimul scenariu a urmărit îmbunătățirea evaluării profilului consumatorilor în ceea ce privește gestionarea eficientă a resurselor de apă, prin crearea unei imagini de ansamblu mai precise a rețelei de distribuție a apei. În comparație cu clasificarea pe gospodării, o abordare bazată pe date poate oferi rezultate mai precise despre tipurile de consumatori în funcție de comportamentul acestora. Metodele de grupare sunt comparate cu modele de clasificare pentru evaluarea profilurilor de consumatori și a gradului de asemănare a acestora cu clasele de consumatori. Abordarea propusă evaluează nivelul de similitudine dintre clasificarea inițială și profilurile consumatorilor obținute din datele de consum efectiv.

2.2.2 Suport decizional pentru gestionarea resurselor de apă (UPB.DSS)

Pentru definirea cazurilor de utilizare am considerat următoarele componente și interacțiuni:

- Data Service - serviciu colectare date de consum;
- ML Service - serviciu procesare date cu algoritmi de învățare automată (machine learning – ML).

Componenta de suport decizional presupune următoarele cazuri de utilizare (Fig. 8):

- Recomandări pe baza caracteristicilor de consum pentru optimizarea costurilor și educarea consumatorului;
- Vizualizare informații despre consumul de apă sub formă de grafice sau reprezentări vizuale pe hartă pentru o imagine de ansamblu asupra întregului sistem și ajustarea recomandărilor pe zone geografice.

Consumatorul și furnizorul interacționează cu sistemul de monitorizare prin intermediul aplicației de monitorizare a consumului de apă (UPB.WMS), la care se adaugă funcțiile de suport decizional reprezentate printr-un set de cazuri de utilizare specifice.

Consumatorul poate să vizualizeze datele de consum și recomandările generate pe baza sistemului de suport decizional. Furnizorul poate vizualiza informații despre consumul de apă sub formă de grafice sau reprezentări vizuale pe hartă, obținute prin prelucrarea avansată a datelor colectate prin algoritmi de învățare automată (clusterizare, clasificare).

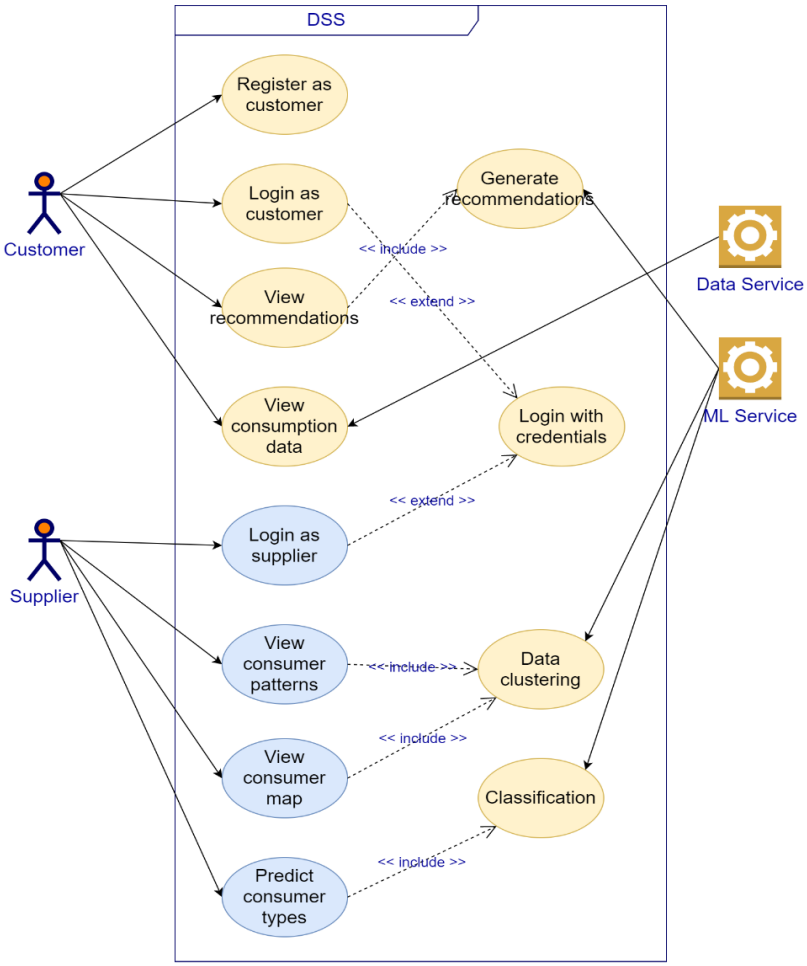


Fig. 8 Diagrama cazurilor de utilizare pentru sistemul de suport decizional

2.3 Sistemul de Mobile Crowdsensing (MCS)

Sistemul colaborativ de raportare prin crowdsensing poate fi descris prin următoarele scenarii generale:

- Abordare complementară sistemelor IoT (Internet of Things);
- Colectarea datelor folosind dispozitive mobile și participarea activă a cetățenilor (crowdsensing, participatory sensing);
- Alocare oportunistă a activităților pe bază de geolocație, nivel de încredere și optimizare a costurilor.

2.3.1 Monitorizarea infrastructurii de apă prin crowdsensing (UPB.CS)

Pentru definirea cazurilor de utilizare am considerat următoarele componente și interacțiuni (Fig. 9):

- Watergame App (aplicație interactivă de raportare incidente);
- User (Blue Team) - utilizator care raportează incidente;
- User (Yellow Team) - utilizator care rezolvă problemele raportate;
- User (Red Team) - utilizator care validează soluțiile;
- Watergame Server - server-ul aplicației.

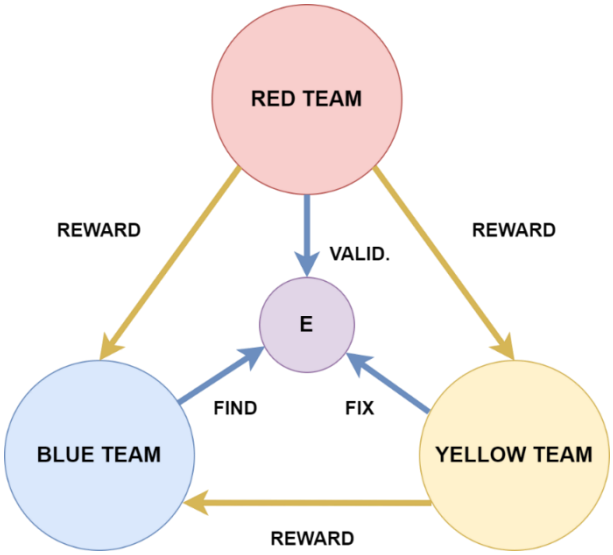


Fig. 9 Diagrama conceptuală a modelului de raportare colaborativă

Implicarea publicului larg este încurajată prin utilizarea elementelor de joc, precum și printr-un sistem de recompense reale. Sistemul are la bază o componentă de joc serios reprezentată printr-un set de roluri / moduri de joc (cazuri de utilizare - Fig. 10):

- **Finder (Blue Team)** – modul pentru utilizatorul general reprezentat de cetățeanul digital care este implicat în comunitate. Utilizatorul deschide harta care prezintă straturile de infrastructură. În cazul în care este identificată o anomalie, utilizatorul poate genera un raport cu geolocalizare.

- **Fixer (Yellow Team)** – modul pentru experți sau utilizatori generali care sunt calificați sau pot să rezolve evenimentele raportate. Utilizatorul deschide harta care arată evenimentele din apropiere și selectează o destinație. Aplicația trece în modul de navigare și arată indicații pentru eveniment. Când evenimentul este confirmat, utilizatorul **Finder** este notificat. În caz contrar, evenimentul poate fi raportat ca fals pozitiv. Sistemul de recompensare poate fi consolidat în continuare prin recompense negative în cazul evenimentelor marcate ca fals pozitiv, astfel încât potențialul de spam să fie redus. A doua sarcină este de a remedia problema și de a încărca o dovadă pentru validare. Când dovada este validată de un alt expert, utilizatorul **Fixer** este notificat.
- **Validator (Red Team)** - modul pentru experți verificați care sunt calificați pentru evaluarea evenimentelor și a problemelor specifice. Sarcina principală a validatorului este evaluarea evenimentelor rezolvate de echipa de teren. Sistemul de recompense în acest caz se bazează pe reputație. Când dovada este validată, utilizatorul **Fixer** este notificat.

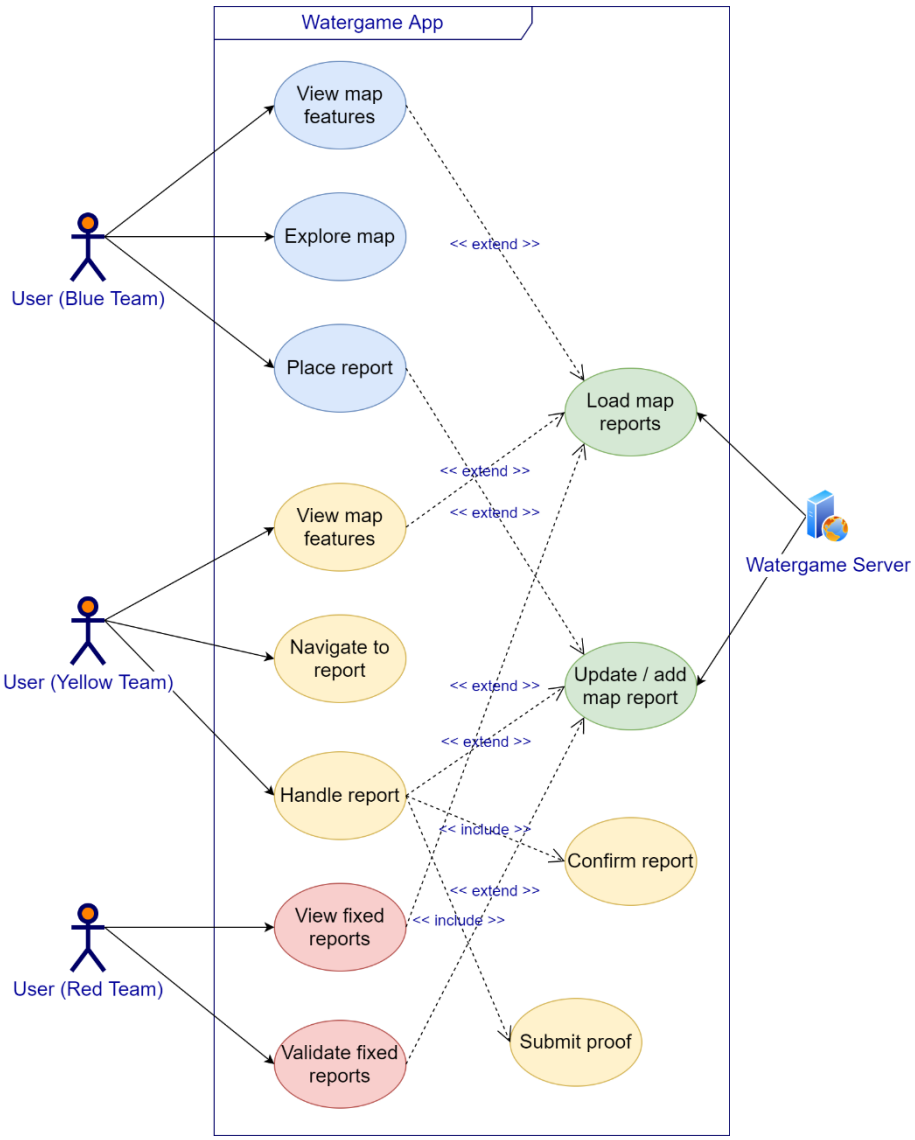


Fig. 10 Diagrama cazurilor de utilizare pentru sistemul colaborativ de raportare

În timp ce primele două cazuri de utilizare (Finder, Fixer) vizează colaborarea și participarea cetățenilor în rezolvarea problemelor de infrastructură în orașe inteligente, al treilea caz de utilizare asigură validarea soluțiilor. Acest comportament este similar într-o oarecare măsură cu tehnologia Blockchain. Interacțiunea dintre participanți și lumea reală este validată de către entități de încredere reprezentate în acest caz de experții verificați (Predescu & Mocanu, 2019; Predescu et al., 2021).

Utilizatorul Finder (Blue Team) explorează harta, vizualizează stratul de infrastructură și poate plasa indicatori pe hartă atunci când identifică un incident în infrastructura de distribuție a apei. Incidentele sunt apoi salvate într-o bază de date sub formă de rapoarte prin intermediul server-ului principal.

Utilizatorul Fixer (Yellow Team) vizualizează rapoartele pe hartă și primește indicații de orientare pentru a se deplasa la locul incidentului. Există posibilitatea de a confirma sau infirma raportul și de a încărca dovada soluționării problemei pentru a fi validată de un expert.

Utilizatorul Validator (Red Team) este un reprezentant al furnizorului de apă, vizualizează și validează rapoartele soluționate de echipa de teren.

2.4 Sistemul Blockchain (B)

Sistemul blockchain poate fi descris prin următoarele scenarii generale:

- Definirea modului de funcționare a soluției blockchain folosind Ethereum (Ethereum-like network);
- Alegerea modului de implementare bazat pe o criptomonedă pe o rețea de blockchain de tip public, în detrimentul unei soluții de tip privat sau a unei soluții bazate pe NFT-uri (Non-fungible token).

2.4.1 Sistem de recompense pe bază de Blockchain (UPB.B)

Pentru definirea cazurilor de utilizare am considerat următoarele componente și interacțiuni (Fig. 11):

- Furnizor (supplier):
  - Administrează contractele inteligente (smart contracts);
  - Inițiază și finalizează ICO-urile (Initial Coin Offering);
  - Acordă puncte pentru raportările de tip crowdsensing;
- Consumator (customer):
  - Cumpără criptomonede;
  - Folosește criptomonede pentru obținere de contracte mai avantajoase;
  - Donează criptomonede.
- Alți utilizatori (other users):
  - Vizualizează numărul de criptomonede pentru diverși utilizatori.

Sistemul permite realizarea acțiunilor pe bază de criptomonedă în afara soluției implementate prin intermediul canalelor de swapping (e.g. pancake swap, uniswap, exchange-uri).

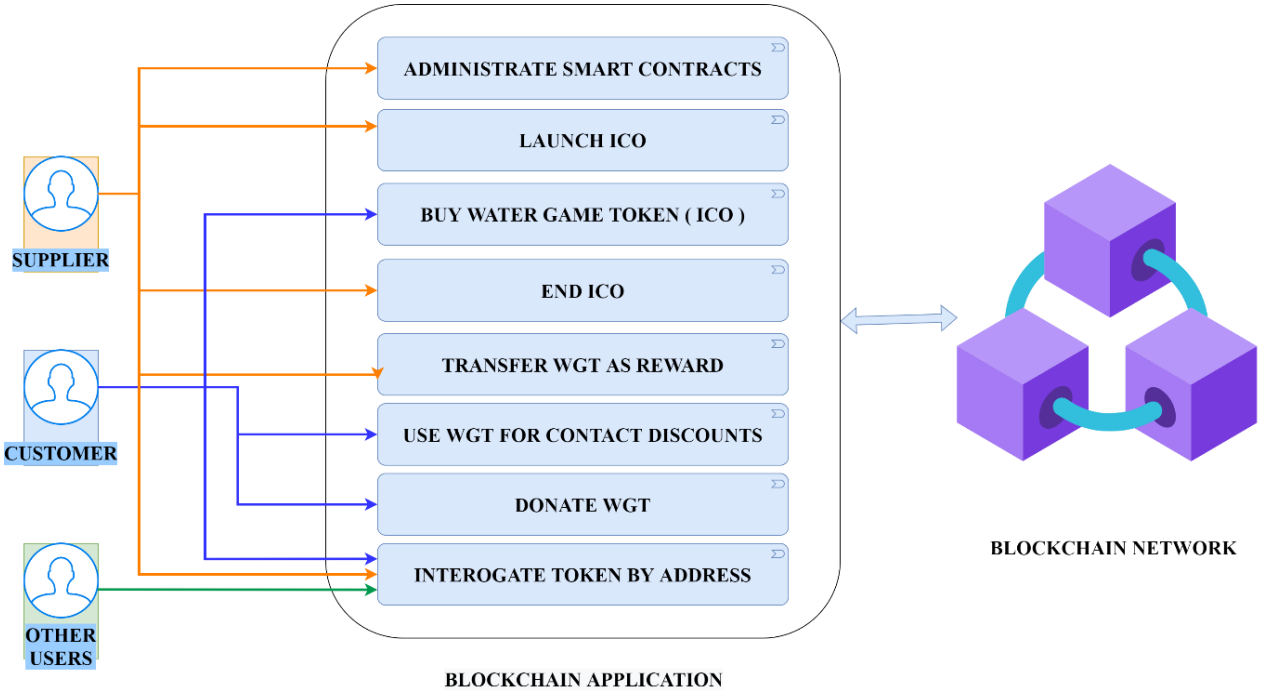


Fig. 11 Diagrama cazurilor de utilizare pentru sistemul de recompense pe bază de blockchain

3 Proiectarea arhitecturii sistemului

Arhitectură generală a sistemului este prezentată în Fig. 12.

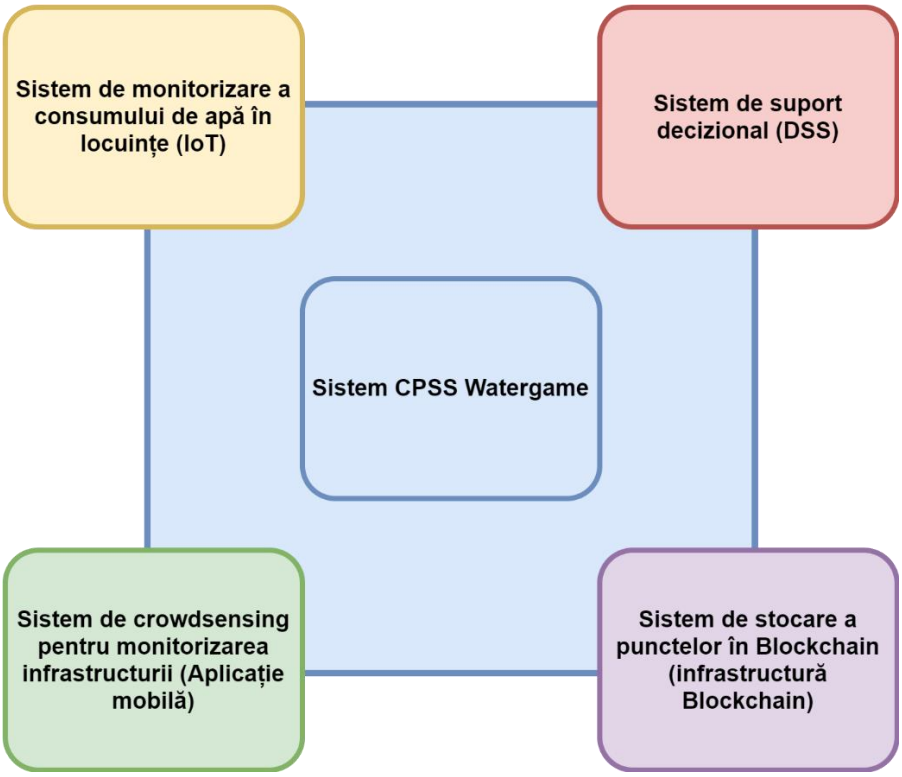


Fig. 12 Arhitectura sistemului CPSS

Sistemul este conceput ca un ansamblu de subsisteme independente, care pot interacționa în contextul unei arhitecturi decuplate. Fiind un sistem complex de tip CPSS, poate fi văzut ca un meta-sistem alcătuit din mai multe componente fundamentale, precum:

- sistemul de monitorizare a consumului de apă în locuințe prin intermediul unei arhitecturi de tip IoT;
- sistemul de suport decizional (DSS) ce oferă funcții inteligente de procesare a datelor;
- sistemul de crowdsensing pentru monitorizarea infrastructurii, prin care este implicat factorul social;
- sistemul de stocare a punctelor în Blockchain.

Integrarea acestor componente este posibilă la nivel de cazuri de utilizare, prin sistemul de puncte ce pot fi obținute din raportări sau prin eficientizarea consumului de apă în gospodărie. Sistemul de suport decizional utilizează datele colectate de la senzori pentru a oferi o imagine de ansamblu pentru furnizori și recomandări personalizate pentru consumatori. Sistemul de crowdsensing are la bază o aplicație mobilă prin care se pot raporta incidente în furnizarea apei sau în contextul infrastructurii de apă la nivel urban.

În contextul acestei etape, sunt prezentate în continuare aceste sub-sisteme din perspectiva arhitecturii și a tehnologiilor folosite, fiecare sub-sistem fiind reprezentat printr-o arhitectură și un set de tehnologii specifice.

3.1 Arhitectura componentei de monitorizare a consumului (WMS)

3.1.1 Arhitectura de colectare și procesare a datelor la UTCN (UTCN.C)

În varianta gândită de UTCN (**UTCN.C**), arhitectura sistemului de colectare și pre-procesare a datelor de consum de apă este de tip multi-nivel, cu 3 nivele (Fig. 13): Monitorizare, Edge și Cloud.

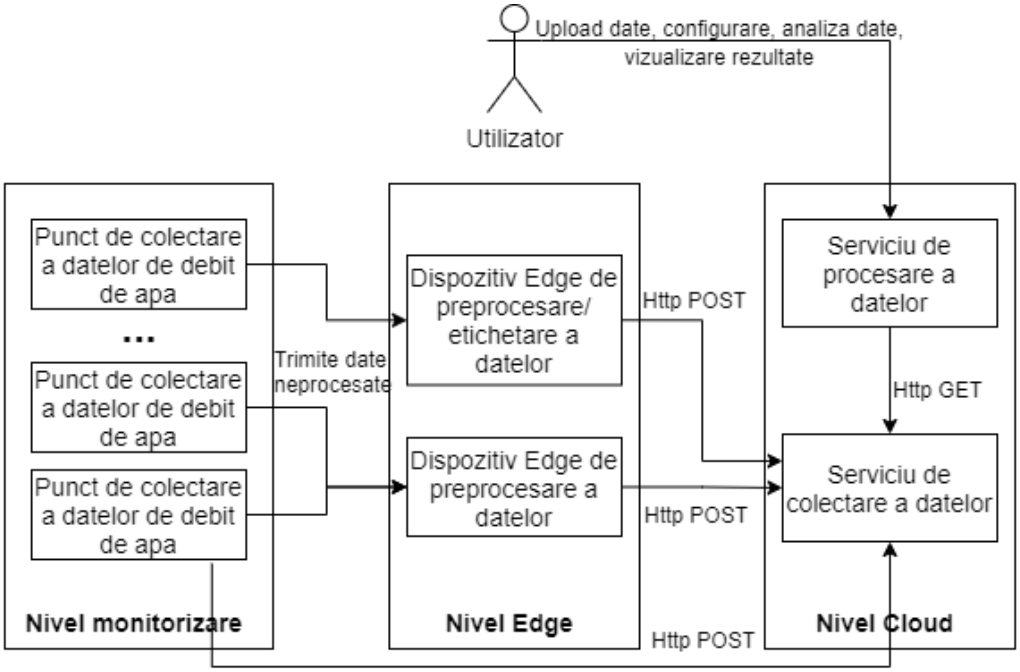


Fig. 13 Arhitectura sistemului de colectare și procesare a datelor de consum de apă (UTCN.C)

La baza arhitecturii se află **nivelul de monitorizare**, care cuprinde punctele în care se colectează datele de debit de apă instalate la locația consumatorului. La o locație pot fi instalate mai multe astfel de puncte de colectare a datelor. Datele de debit, neprocesate, se trimit periodic, în batch-uri dispozitivelor de la nivelul edge.

Nivelul edge este compus din dispozitive cu capacitate de stocare și procesare redusă, care aplică proceduri de pre-procesare a datelor primite de la punctele de colectare a datelor de debit. Datele pre-procesate sunt trimise apoi serviciului de colectare a datelor de la nivelul cloud. La nivelul edge sunt păstrate datele dintr-o fereastră de timp configurată în funcție de capacitatea de stocare a dispozitivului folosit.

La **nivelul cloud** se află un serviciu de colectare, stocare de lungă durată a datelor și un serviciu de procesare avansată (offline) a datelor. Serviciul de colectare primește date preprocesate de la dispozitivele edge și le stochează într-o bază de date în format standard OGC SensorThings[3] sau poate primi cereri prin care să servească datele stocate altor aplicații.

Serviciul de procesare a datelor poate aplica proceduri complexe de preprocesare și analiză, folosind algoritmi de învățare automată pe datele stocate de serviciul de colectare a datelor de consum sau pe seturi de date provenite din alte surse. Serviciul de procesare a datelor are atașată

[3] https://www.ogc.org/standards/sensorthings

o interfață utilizator care permite utilizatorului să încarce seturi de date, să configureze și să execute pipeline-uri de procesare a datelor și să vizualizeze rezultatele analizei datelor.

3.1.2 Arhitectura de colectare a datelor la UPB (UPB.C)

Varianta UPB pentru colectarea eficientă a datelor legate de consumul de apă (**UPB.C**) descrisă în Fig. 14 presupune că rețeaua de senzori este integrată într-o arhitectură bazată pe servicii Cloud ce permite evitarea nivelului edge din sistemul de colectare gândit de UTCN (Fig. 13).

Astfel, achiziția datelor se realizează cu ajutorul unei plăci de dezvoltare NodeMCU bazată pe microcontrolerul ESP8266 care dispune de comunicare WiFi, pini GPIO (General-purpose input / output) și poate fi programat în mediul Arduino (Parihar, 2019). O astfel de placă de dezvoltare poate fi conectată la mai multe debitmetre, monitorizând astfel debitul de apă prin mai multe conducte. Modulul de achiziție astfel proiectat, este pre-programat și dispune de o interfață de configurare ce permite conectarea la rețeaua locală și atașarea unui număr variabil de debitmetre. Interfața este expusă printr-un server găzduit direct pe microcontrolerul ESP8266 configurat în același timp în modul stație și client WiFi.

Colectarea datelor se realizează printr-un broker de mesaje MQTT, ce permite o arhitectură decuplată printr-un protocol standard de comunicație, adaptat pentru domeniul IoT (Mishra & Kertesz, 2020). MQTT este un protocol de comunicație de tip publish-subscribe ce suportă conexiune bi-direcțională peste TCP / IP, fiind ideal pentru schimbul de mesaje între dispozitive IoT (IBM, 2021). Dispozitivele IoT se pot astfel conecta la broker-ul de mesaje ce are rol de rutare a mesajelor primite de la clienți, și pot publica mesaje pe un topic sau se pot abona la un topic pentru a primi mesaje.

Server-ul aplicației se conectează la broker-ul MQTT pentru a recepționa datele trimise de senzori și realizează interfața cu baza de date MySQL[4] (Bell, 2016). Datele colectate sunt salvate periodic în baza de date, conform cu intervalul configurat la nivel de dispozitiv IoT. Există două topice pentru transmiterea datelor: date de consum instantaneu (monitorizare în timp real, cu frecvență ridicată), date de volum (înregistrate în baza de date, cu o frecvență mai scăzută). Datele pot fi accesate în timp real folosind un client MQTT (e.g. aplicație desktop, module IoT), conectarea fiind permisă pe bază de credențiale de acces la nivel de broker.

Pentru disponibilitatea datelor pentru vizualizare în timp real, este adăugat un modul REDIS[5] pentru stocarea temporară a ultimelor măsurători colectate pe topicul de consum instantaneu (Chen et al., 2016).

Datele legate de consum și starea senzorilor sunt prezentate către client cu ajutorul unei aplicații mobile (Android). Autentificarea și gestionarea conturilor pentru utilizatori se realizează prin serviciul Firebase[6], iar încărcarea datelor de consum se realizează prin cereri HTTP de la server-ul principal.

[4] https://www.mysql.com/

[5] https://redis.io/

[6] https://firebase.google.com/

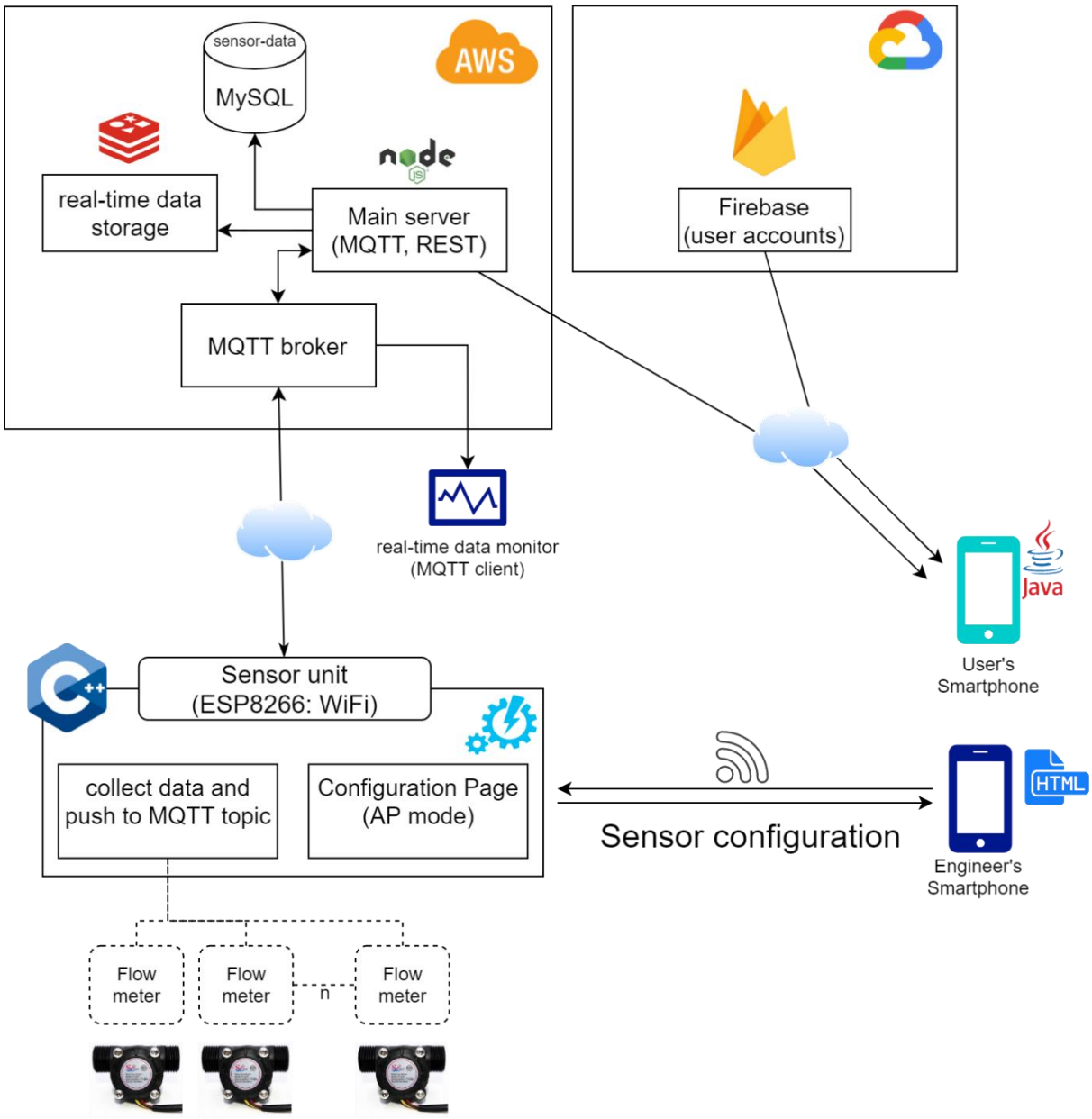


Fig. 14 Arhitectura sistemului de monitorizare a consumului de apă în locuințe (UPB.C)

3.2 Arhitectura componentei de suport decizional (DSS)

Componenta de suport decizional reprezintă o extensie a sistemului de monitorizare a consumului, realizând analiza datelor și apoi oferind rapoarte și recomandări pentru furnizori sau consumatori, în vederea eficientizării consumului de apă.

Sistemul de tip DSS integrează algoritmi de ML într-un serviciu implementat în Python care preia datele de consum de la server-ul aplicației și implementează scenariile de procesare a datelor. Aplicația mobilă prezintă recomandările individuale generate în urma procesării pentru consumator.

3.3 Arhitectura componentei de Crowdsensing (CS)

Arhitectura componentei de crowdsensing (Fig. 15) are la bază servicii Cloud și o aplicație mobilă interactivă. Pentru a asigura infrastructura tehnică necesară unei aplicații la standardele actuale din industrie, sunt considerate integrări cu servicii de aplicație precum: Firebase[7], Google Sign In, OneSignal[8], Apple Login.

Spre deosebire de celelalte componente, factorul uman este central în funcționarea aplicației pentru raportarea incidentelor pe hartă. Aplicația permite plasarea de indicatori pe hartă în mod interactiv, pe bază de geolocație și realitate mixtă, combinate cu elemente de joc pentru stimularea participării.

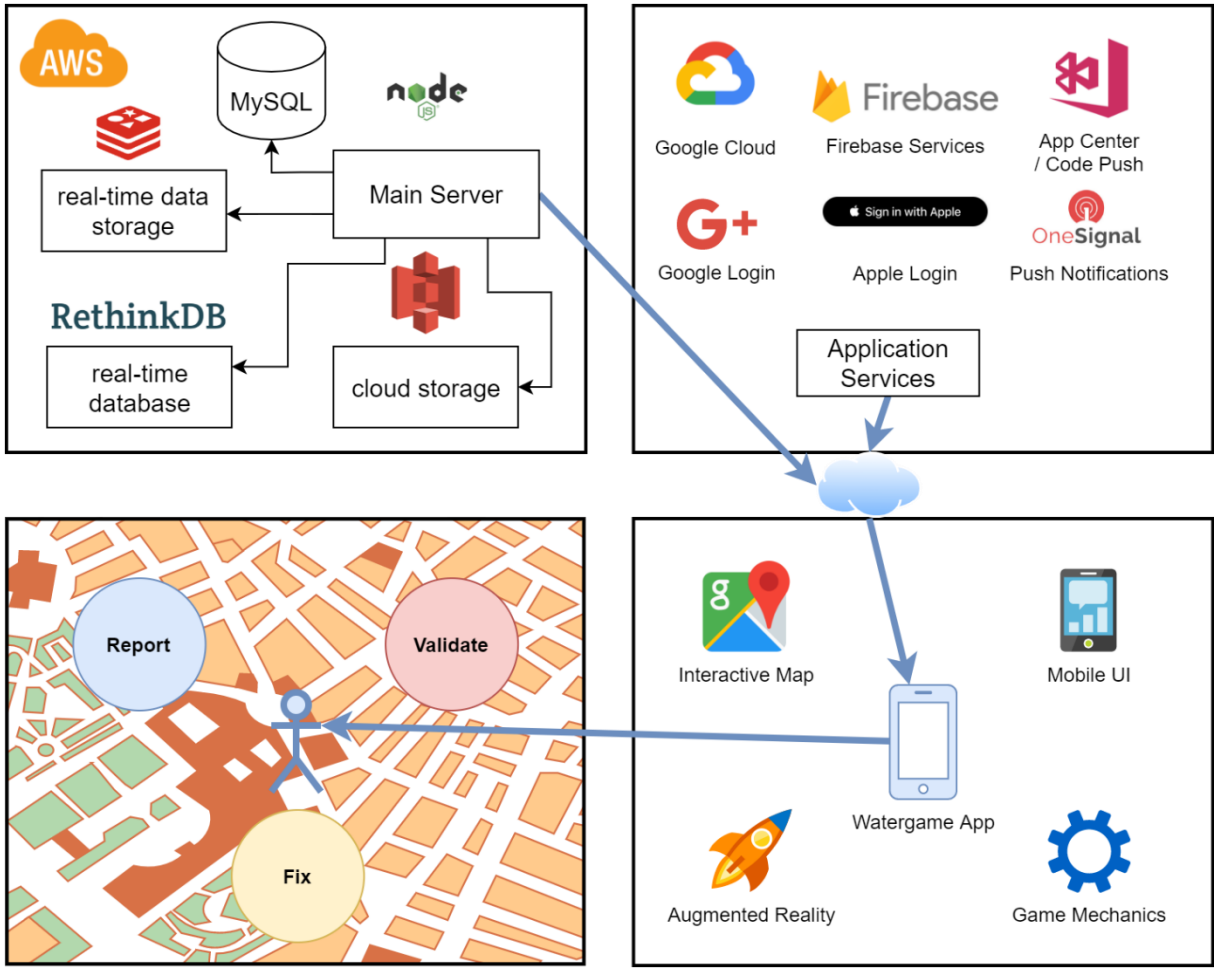


Fig. 15 Arhitectura sistemului de raportare a incidentelor pe bază de crowdsensing

[7] https://firebase.google.com/

[8] https://onesignal.com/

Pentru operabilitate în condiții reale sunt considerate metode variate de stimulare a participării precum:

- modele matematice pentru alocarea dinamică a activităților de raportare;
- gamificare și elemente de joc serios (Serious Gaming);
- aspect vizual atractiv și interfață intuitivă.

Infrastructura Cloud este găzduită pe AWS (Amazon Web Services) reprezentată printr-un server de aplicație care gestionează cererile HTTP de la aplicația mobilă precum și logica de business pentru accesarea bazei de date și a sistemelor de stocare a datelor temporare. Sistemul operează cu o bază de date MySQL ce asigură persistența datelor și o structură bine definită pentru operabilitate. Pentru scalabilitate, sistemul incorporează și o memorie temporară de tip REDIS. Datele în timp real sunt procesate folosind o bază de date de timp real RethinkDB[9] ce permite funcții avansate de interogare a datelor (Wingerath, 2017) în timp ce datele multimedia sunt stocate într-un serviciu de stocare cloud (Amazon S3).

Arhitectura jocului este reprezentată de cele 3 scenarii specifice fiecărui tip de utilizator (Finder, Fixer, Validator) și are la bază harta interactivă configurată pentru a integra elementele de joc propuse.

3.4 Arhitectura componentei Blockchain (B)

Arhitectura componentei Blockchain (Fig. 16) are la bază o rețea de test Ethereum și o aplicație web funcțională în condiții de laborator. Interacțiunea cu rețeaua blockchain se realizează prin intermediul aplicației sau prin intermediul unui wallet conectat direct (e.g. MetaMask (MetaMask, 2021)). Pentru a sincroniza wallet-urile cu conturile de utilizator din sistem, sistemul va fi conectat indirect (prin server-ul de aplicație, folosind cereri HTTP) la baza de date MySQL de pe AWS.

Rețeaua de test este pusă la dispoziție de platforma Ethereum și presupune o serie de noduri conectate între ele în mod peer-top-peer, care oferă posibilitatea introducerii datelor în blockchain, validării tranzacțiilor și interogării datelor (Neto, 2020).

Aplicația web prin intermediul căreia se va putea tranzacționa token-ul (WTG – Water Game Token) va fi instalată folosind NPM Lite Server[10]. Soluția va oferi și diverse validări pentru operațiunile ce se doresc a fi efectuate (donare de token, transfer token) prin intermediul JavaScript și web3.js. De asemenea, prin intermediul Web3.js se va realiza conexiunea la blockchain și interacțiunea cu Smart Contracts (Ethereum Revision a57dd3c5, 2016).

Pentru efectuarea diferitelor operațiuni pe blockchain (e.g. oferirea de token drept recompensă pentru utilizatori), ce vor fi declanșate din exteriorul rețelei, se va pune la dispoziție un REST (Representational state transfer) API prin intermediul căruia se vor realiza respectivele cereri.

[9] https://rethinkdb.com/

[10] https://www.npmjs.com/package/lite-server

Arhitectura soluției va oferi posibilitatea tranzacționării token-ului generat și din afara soluției, prin intermediul alternativelor precum: exchange-uri sau platforme de tranzacționare de tokeni, dacă se dorește trecerea monedei într-un mediu de producție.

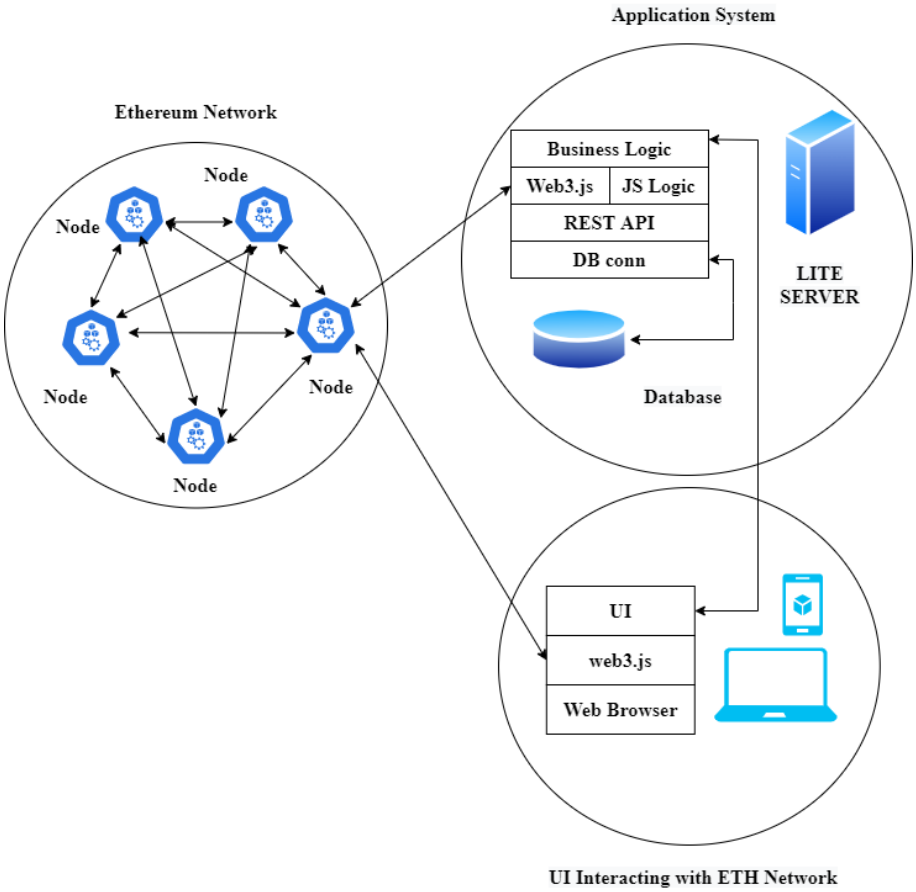


Fig. 16 Arhitectura componentei blockchain

4 Analiza datelor

4.1 Seturi de date existente

4.1.1 Helix

Pentru identificarea și modelarea corespunzătoare a datelor relevante pentru a fi colectate, au fost căutate seturi de date publice similare. Astfel, a fost identificat un set de date legat de consumul de apă din orașul Alicante, denumit Helix (Chatzigeorgakidis, 2018). În orașul Alicante principalul furnizor de apă este Aguas Municipalizadas de Alicante, Empresa Mixta (AMAEM). În trecut, consumul de apă a fost monitorizat prin citirea directă a contoarelor și facturarea a fost procesată la fiecare trei luni. Începând cu 2013, AMAEM a schimbat toate contoarele, trecând de la contoare de apă analogice la cele inteligente (IWM), care înregistrează automat datele. Mai mult, informațiile au fost stocate electronic și transmise furnizorului printr-o rețea radio. În mod normal, IWM-urile pot citi date la fiecare câteva secunde, dar pentru o durată mai mare de viață a bateriei (~10 ani), datele au fost citite din ora în ora. Drept urmare, acest set de date analizat conține informații de la 822 de gospodării din orașul Alicante, colectate pe durata a 2 ani, începând cu 01.01.2015 și terminând pe 20.01.2017.

Fiecare intrare din setul de date al gospodăriei conține următoarele informații:

- un identificator unic și anonimizat al gospodăriei;
- detalii despre poziția geografică a gospodăriei (latitudine geografică carteziană și longitudinea locuinței);
- consumul mediu pentru fiecare oră pe săptămână (pentru un total de 24 × 7 = 168 valori săptămânale).

4.1.2 GECCO

Un alt set de date identificat a fost GECCO 2017 (Chandrasekaran et al., 2017). Acest set de date conține peste 120.000 de exemple de date reale despre apa potabilă. Conține 9 atribute măsurate și o coloană țintă, o valoare binară reprezentând dacă există sau nu o schimbare remarcabilă în calitatea apei. Atributele măsurate sunt următoarele:

- Tp - reprezintă temperatura apei exprimată în grade Celsius;
- pH - valoarea pH-ului apei;
- Leit - conductivitate electrică;
- Trueb - turbiditatea apei;
- Redox - potențial redox;
- Cl, Cl2 - dioxid de clor;
- Fm, Fm2 – debit.

Valorile prezentate mai sus sunt indicatori foarte buni de calitate a apei. De exemplu, cantitatea de solide dizolvate în apă determină conductivitatea electrică, deci acesta este un bun indicator al calității apei. Turbiditatea este un indicator al transparenței relative a lichidului. Argila, nămolul, substanțele anorganice foarte mici și materia organică, cum ar fi microorganismele, pot fi

detectate pe baza turbidității apei. Vorbind despre pH-ul apei de băut, în general, intervalul normal pentru pH este de la 6,5 la 8,5.

4.1.3 Austin Water

Ultimul set de date identificat este Austin Water (2021). Setul conține consumul lunar de apă al consumatorilor, fiind grupat în funcție de codul poștal și categoria de consumator. Colectarea datelor a durat aproximativ 5 ani, începând din ianuarie 2012 și terminându-se în octombrie 2017. Setul de date este format din următoarele câmpuri:

- Cod_An_Lună - reprezintă o concatenare între an și lună –(de ex.: 201201);
- Postal_Code - codul poștal în funcție de care se identifica locația consumatorului (pot exista mai multe tipuri de consumatori pentru fiecare cod poștal);
- Client_Class - au fost definite patru tipuri de clase de consumatori:
  - Multifamilial;
  - Irigare – Multifamilial;
  - Rezidențial;
  - Irigare – Rezidențial;
- Total_Galoane - reprezintă numărul total de litri consumați de fiecare din cele patru clase prezentate mai sus, într-o luna, grupate după codul poștal.

4.2 Scenarii de prelucrare a datelor

Scenariile de prelucrare a datelor au fost implementate pe baza datelor disponibile anterior, ca un set de script-uri Python[11], în vederea identificării caracteristicilor necesare pentru colectarea datelor în cadrul sistemului proiectat.

4.2.1 Detecția anomaliilor (Scenariul UTCN.A1)

O provocare majoră în sistemele de alimentare cu apă este identificarea automată a oricărei abateri de la o stare considerată normală sau, cu alte cuvinte, detecția anomaliilor. Un comportament anormal poate fi rezultatul mai multor factori: modificări ale surselor de apă (de exemplu, debitul, calitatea etc.), defecte ale infrastructurii de distribuție (de exemplu, senzori sau actuatoare defecte, scurgeri de conducte) sau atacuri rău intenționate. Efectele unei anomalii în sistemul de apă pot influența sănătatea unei părți semnificative a populației. Prin urmare, orice incident (funcționare defectuoasă) trebuie detectat în cel mai scurt timp posibil și măsurile de recuperare trebuie inițiate, dacă este posibil, într-un mod automat.

Pentru a crea și a antrena un model capabil să detecteze anomaliile în calitatea apei potabile am folosit setul de date numit GECCO 2017 (Chandrasekaran et al., 2017).

Scopul principal în acest caz este detectarea modificărilor / evenimentelor din setul de date. Deoarece aceste modificări sunt rare, setul de date este foarte dezechilibrat (vezi Fig. 17), astfel încât problema poate fi considerată o problemă de detecție a anomaliilor.

[11] https://github.com/alexp25/watergame-other

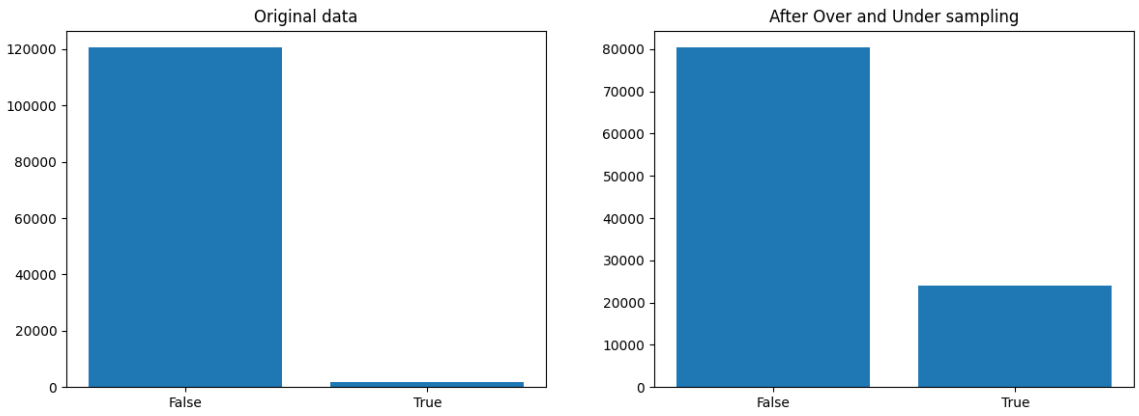


Fig. 17 Setul de date original și re-echilibrat folosind SMOTE și RUID

Pentru a crește performanța modelelor inteligență artificială (AI) / ML, este o bună practică să preprocesăm datele brute înainte de a le furniza unui algoritm de clasificare. Zgomotul, datele lipsă sau datele dezechilibrate pot afecta negativ performanța modelului de AI.

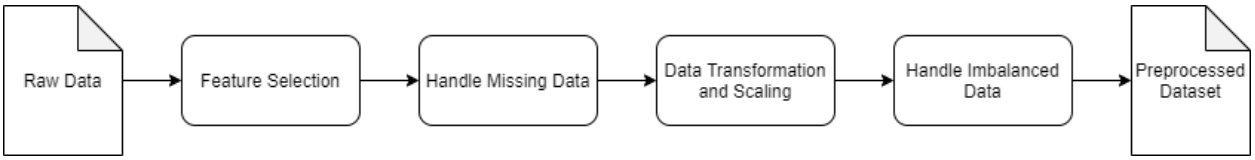


Fig. 18 Pipeline de pre-procesare a datelor GECCO

Fig. 18 prezintă pipeline-ul de preprocesare a datelor folosit pentru a pregăti setul de date privind calitatea apei potabile. Deoarece există aproximativ 10.000 de valori lipsă (în medie) în fiecare coloană a setului de date, este absolut necesar să gestionăm datele lipsă. Am încercat mai multe abordări: în primul rând, am eliminat rândurile care conțineau aceste valori lipsă, dar în acest fel am pierdut prea multe exemple valoroase. Următoarea încercare a fost să înlocuim valorile lipsă cu valoarea medie, mediană sau cu zerouri. Soluția cea mai potrivită a fost completarea coloanelor lipsă cu zerouri. Acest lucru a dus la cea mai mare performanță a modelului.

Există mai mulți algoritmi precum SVM sau rețele neuronale, la care scări diferite ale datelor pentru diferite coloane pot afecta negativ performanța modelului. Pentru a depăși această problemă, am transformat datele folosind diferite strategii de transformare, cum ar fi scalarea min-max, scalarea standard, scalarea în intervalul [0, 1] și normalizarea. Cele mai bune rezultate au fost date de scalarea în intervalul [0, 1] sau scalarea standard.

În ultimul pas, deoarece datele de intrare sunt foarte dezechilibrate, am folosit SMOTE (Chawla et al., 2002) pentru a supra-eșantiona clasa minoritară și RUID (Prusa et al., 2015) pentru sub-eșantionarea clasei majoritare, încercând astfel să reechilibrăm setul de date. După rularea acestor algoritmi, am obținut un set de date cu un număr mai apropiat de etichete (vezi Fig. 17)

Tabel 1 Rezultatele diferiților algoritmi AI folosind setul de date GECCO

ALGORITHMS PERFORMANCE

Algorithm	*F1-Measure*
RANDOM FOREST	99.93%
EXTRA TREES CLASSIFIER	99.81%
DECISION TREE	99.79%
MLP	99.49%
KNN	99.47%
One-Class SVM	81.46%
SGD CLASSIFIER	50.36%
LOGISTIC REGRESSION	49.66%
PASSIVE AGRESSIVE CLASSIFIER	45.36%
RIDGE CLASSIFIER	37.34%

The best algorithm is RANDOM FOREST with 99.93%.
Tuned parameters:
class_weight='balanced', max_depth=42, n_estimators=130

După antrenarea mai multor algoritmi de AI (așa cum se poate vedea în Tabel 1), modelul cel mai eficient s-a dovedit a fi Random Forest (Breiman, 2001) cu o măsură F1 de testare de 99,93% și o măsură F1 de validare (folosind date nemaivăzute) de 99,49%. Matricea de confuzie folosind datele de validare este prezentată în Fig. 19.

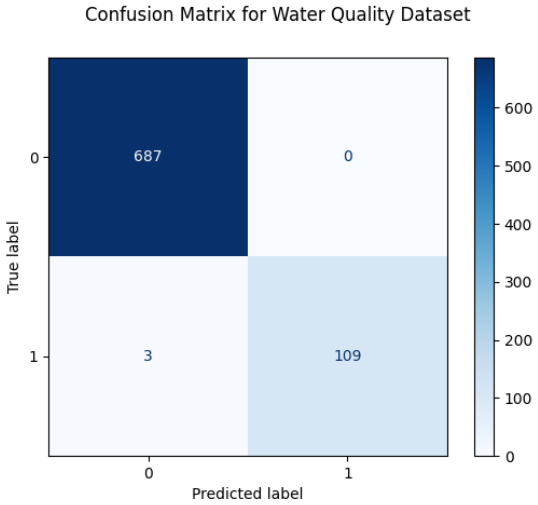


Fig. 19 Matricea de confuzie folosind date de validare

Există mai mulți algoritmi care au obținut rezultate similare, cum ar fi Random Forest, Extra Tree classifier (Geurts, Ernst, & Wehenkel, 2006) sau Decision Tree (Topîrceanu & Grosseck, 2017). Acest lucru este normal, deoarece acești algoritmi sunt foarte similari, toți sunt metode de ansamblu bazate pe arbori. Multi Layer Perceptron (Murtagh, 1991) a obținut, de asemenea, un scor F1 ridicat. Acest scor a fost obținut prin utilizarea a 9 straturi complet conectate, începând cu 512 noduri și apoi înjumătățind numărul de noduri la fiecare strat următor, având în final 2 noduri pentru a clasifica două clase. Chiar dacă MLP Classifier a obținut un scor F1 mare, timpul de antrenament este mult mai mare decât în cazul Random Forest sau Extra Tree Classifier, așa că, în general, câștigătorul acestei curse a fost Random Forest.

Pentru a explica și înțelege rezultatele modelului selectat, am folosit importanța caracteristicilor care indică importanța relativă a fiecărei caracteristici atunci când se face o predicție

(vezi în Fig. 20). După cum putem observa în această figură, cea mai importantă caracteristică (sau atribut) pentru detectarea anomaliilor în calitatea apei potabile este Redox, sau cu alte cuvinte, potențialul de reducere de oxidare, care indică cât de igienizată sau contaminată este apa pe baza oxidării și reducerii acesteia. Aceasta este o măsură foarte bună pentru estimarea calității apei, așa că orice sistem de monitorizare a calității apei ar trebui să utilizeze contoare ORP.

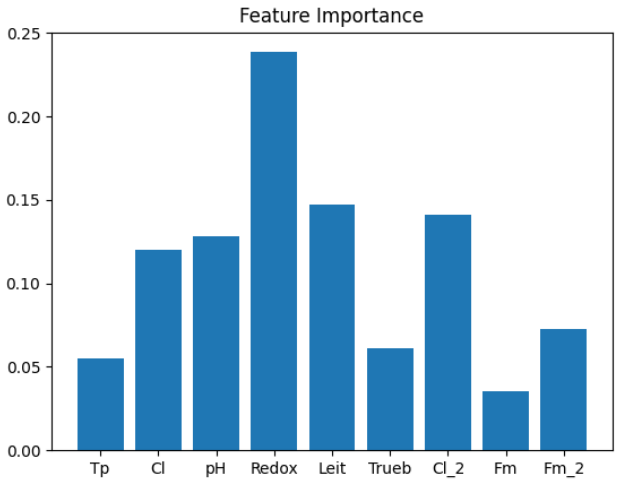


Fig. 20 Diagrama de importanță a caracteristicilor

Pentru a doua poziție în importanța caracteristicii, avem mai multe atribute, care au valori de importanță foarte apropiate, cum ar fi pH, Cl, Cl_2 și Leit. Am fi putut ghici acest rezultat deoarece, de exemplu, dioxidul de clor (Cl și Cl_2) este folosit ca dezinfectant și pentru controlul mirosului și gustului în igienizarea alimentelor, deci acesta poate fi un bun indicator al calității apei. Nivelul pH-ului este, de asemenea, un bun indicator al calității apei, deoarece apa potabilă ar trebui să aibă o valoare a pH-ului între 6,5 și 8,5. Un nivel de pH foarte scăzut sau ridicat poate indica o poluare chimică a apei. Atributul numit Leit este conductivitatea electrică a apei. Cantitatea de solide dizolvate în apă determină conductivitatea electrică, deci acesta este și un bun indicator al calității apei.

Cele mai puțin importante caracteristici sunt atributele numite FM și FM_2. Aceste două atribute indică debitul, care de fapt nu este un indicator al calității apei potabile, ci este un indicator al cantității de apă care iese de la robinet într-un anumit interval de timp. Turbiditatea apei are, de asemenea, o valoare scăzută a importanței caracteristicilor, ceea ce poate fi un semn că acest atribut poate fi redundant și poate fi exprimat prin combinarea altor atribute din setul de date.

4.2.2 Predicția consumului de apă (Scenariul UTCN.A2)

Setul de date Helix (Chatzigeorgakidis, 2018) a fost folosit ca o problemă de predicție a seriilor de timp multi-variate și în mai multe etape / mai mulți pași. Pentru a rezolva această problemă am încercat mai multe soluții, ca, de exemplu Auto-ARIMA (Yermal & Balasubramanian, 2017) sau LSTM (Lu et al., 2020).

Auto-Arima

Descrierea algoritmului

ARIMA este o tehnică foarte populară pentru modelarea seriilor de timp. Ea descrie corelația dintre punctele de date și ia în considerare diferența dintre valori. Acest algoritm funcționează pe următoarele ipoteze/presupuneri:

- Seria de timp este staționară, ceea ce înseamnă că media și varianța nu ar trebui să varieze în timp. O serie poate fi făcută staționară folosind metoda numită log transformation sau prin diferențierea seriei.
- Datele furnizate ca intrare trebuie să fie o serie uni-variată, deoarece ARIMA folosește valorile trecute pentru a prezice valorile viitoare.

ARIMA are trei componente – AR (termen autoregresiv), I (termen de diferențiere) și MA (termen mediu mobil):

- Termenul AR se referă la valorile trecute utilizate pentru prognoza următoarei valori. Termenul AR este definit de parametrul ”p” în algoritmul Arima.
- Termenul MA este folosit pentru a defini numărul de erori de prognoză trecute utilizate pentru a prezice valorile viitoare. Parametrul ”q” în Arima reprezintă termenul MA.
- Ordinea de diferențiere specifică de câte ori este efectuată operația de diferențiere pe serie pentru a o face staționară. Acest termen este definit de parametrul ”d”.

Deși ARIMA este un model foarte puternic pentru prognoza datelor din serii de timp, procesele de pregătire a datelor și de reglare a parametrilor ajung să fie foarte mari consumatoare de timp. Înainte de a implementa ARIMA, trebuie să se facă seria staționară și să se determine valorile lui p și q. Auto-ARIMA ne ajută cu setarea parametrilor și pregătirea datelor de intrare, deci e mult mai ușor de folosit și necesită mai puțin timp pentru implementare și utilizare. De aceea, în acest proiect am ales să utilizăm Auto-ARIMA pentru predicția valorilor de consum de apă. Această soluție poate fi folosită pe dispozitivele Edge, deoarece nu necesită multă memorie și timpul de antrenare e foarte scurt, astfel prin Auto-ARIMA putem genera predicții ale consumului de apă specifice unei gospodării, aproape instant, pe dispozitivul instalat în locuința respectivă, offline, fără să mai trimitem datele către un server central sau în cloud.

După rularea algoritmului Auto-Arima pe prima serie de timp din setul de date Helix (Chatzigeorgakidis, 2018), am obținut modelul final ARIMA(5, 0, 1), unde valorile reprezintă valorile optime găsite pentru parametrii p, q și d.

Evaluarea modelului

Pentru evaluarea și selecția modelului am folosit metrica AIC (Akaike information criterion) (Portet, 2020). AIC este deosebit de valoros pentru seriile de timp, deoarece datele cele mai valoroase ale analizei seriilor de timp sunt adesea cele mai recente, care sunt blocate în seturile de validare și de testare. Ca rezultat, antrenarea pe toate datele și utilizarea AIC poate duce la o selecție îmbunătățită a modelului față de metodele tradiționale de selecție a modelului de antrenare / validare / test.

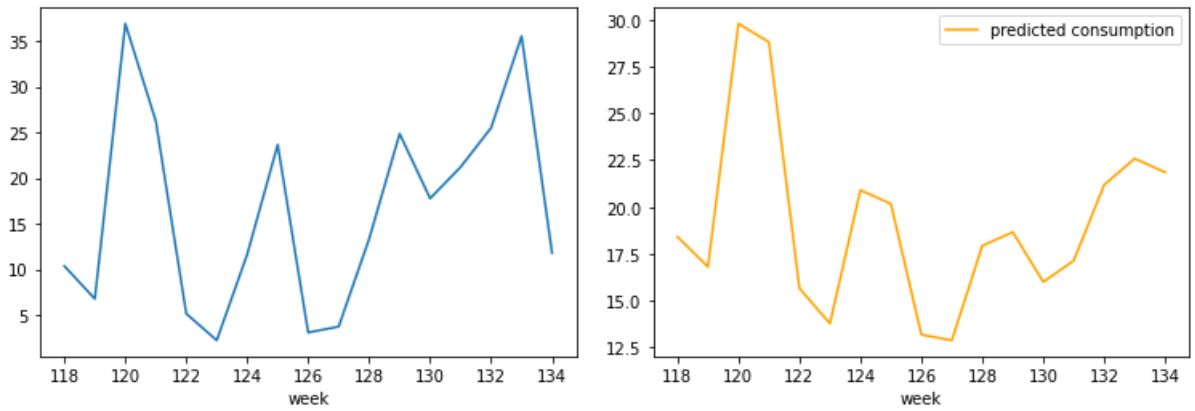


Fig. 21 Rezultatele aplicării Auto-ARIMA pe setul de date Helix (Chatzigeorgakidis, 2018)

Rezultate

Modelul găsit prin Auto-ARIMA are un AIC de 935.49, care e o valoare destul de bună, dacă vrem să obținem trend-ul de mișcare a valorilor și nu valorile exacte. În Fig. 21 putem vedea valorile prezise de modelul antrenat și valorile reale. Culoarea albastră reprezintă valorile reale, iar culoarea portocalie reprezintă valorile prezise. Așa cum putem observa din aceste grafice, rezultatele prezise sunt foarte aproape de realitate; de fapt, tendința de consum de apă este urmărită (forma graficului prezis și forma graficului real este similară), doar valorile nu sunt perfect identice, dar de cele mai multe ori sunt foarte apropiate.

LSTM

Descrierea algoritmului

Ideea din spatele Rețele Neuronale Recurente (Recurent Neural Networks - RNN) (Sherstinsky, 2020) este de a folosi informații secvențiale. Într-o rețea neuronală tradițională presupunem că toate intrările (și ieșirile) sunt independente unele de altele. Dar pentru multe sarcini, această idee nu are rezultate bune. Dacă vrem să prezicem următoarea valoare dintr-o serie de timp, de obicei trebuie să ne bazăm și pe valorile anterioare; deci, în acest caz, nu putem considera că intrările sunt independente. RNN-urile sunt numite recurente deoarece efectuează aceeași sarcină pentru fiecare element al unei secvențe, rezultatul fiind dependent de calculele anterioare și au o „memorie” care captează informații despre ceea ce a fost calculat până acum.

Rețelele neuronale recurente suferă de memorie pe termen scurt, adică dacă o secvență este suficient de lungă, acest tip de rețea neuronală va avea dificultăți în transportarea informațiilor de la pașii de timp anteriori la pașii de mai târziu. Rețelele numite Long Short-Term Memory (LSTM) sunt un tip de rețea neuronală recurentă capabilă să învețe dependența de ordine în problemele de predicție a seriilor de timp, având o natură ce reține informații pentru perioade lungi de timp. Acestea au apărut pentru a rezolva problema principală a RNN-urilor.

LSTM-urile au mecanisme interne numite porți care pot regla fluxul de informații. Aceste porți pot afla care date dintr-o secvență sunt importante pentru a le păstra sau a le arunca / șterge din memorie, astfel se pot transmite informații relevante în lanțul lung de secvențe pentru a face predicții. Aproape toate rezultatele de ultimă generație bazate pe rețele neuronale recurente sunt obținute cu aceste rețele.

Transformarea datelor

Deoarece avem mai multe locuințe și pentru fiecare locuință vrem să prezicem valorile de consum pentru perioada imediat următoare, putem vorbi de tip prognoză sau predicție multi-variată. Un model LSTM se așteaptă ca datele să aibă forma (eșantioane, intervale de timp, caracteristici), deci în acest caz forma lui X este tridimensională, incluzând numărul de eșantioane, numărul de pași de timp aleși per eșantion și numărul de serii de timp paralele sau caracteristici. În cazul nostru concret, deoarece avem 822 de gospodării, vom avea 822 de caracteristici, fiecare având 168 de eșantioane. Presupunând că vrem să prezicem consumul pentru o săptămână întreagă, vom avea 7 zile, deci vom avea date de intrare de forma (24, 7, 822), din care 20 de eșantioane au fost folosite pentru antrenare, adică dimensiunea datelor de antrenare a fost (18, 7, 822) iar restul au fost folosite pentru testare și evaluare, adică dimensiunea datelor de antrenare a fost (6, 7, 822).

Pentru a avea rezultate cât mai corecte, în primul pas valorile lipsă au fost înlocuite cu valoarea din ziua anterioară, astfel garantând că aceste valori lipsă nu afectează predicția finală. În pasul următor s-a calculat transpusa matricei de intrare, deoarece valorile de consum pentru diferite gospodării erau pe linii, iar pentru LSTM, acesta valori (caracteristici) trebuie să fie pe coloane, fiecare coloană reprezentând consumul de apă pentru o gospodărie.

Evaluarea modelului

Atât Root Mean Squared Error (RMSE) cât și Mean Absolute Error (MAE) se potrivesc cu problema evaluării seriilor de timp, deși RMSE este mai frecvent utilizat, așa că și noi am preferat RMSE, deoarece spre deosebire de MAE, RMSE pedepsește mai mult erorile de prognoză. Am calculat valoarea RMSE pentru fiecare valoare prezisă separat dar și împreună, deoarece un singur scor, un scor rezumat e foarte util când trebuie să selectăm modelul cel mai performant dintre mai multe modele.

Rezultate

Acest model, care conține 3 straturi de LSTM cu câte 200 de neuroni pe fiecare nivel, iar la final un strat complet conectat de 1000 de neuroni și unul de 822 pentru cele 822 de gospodării pentru care vrem să facem predicție, a fost antrenat pentru 500 de epoci, iar RMSE-ul final, obținut pe datele de test este de 6.72, care este o eroare foarte mică, având în vedere numărul mare de gospodării pentru care se fac predicții în mod simultan.

În diagramele din Fig. 22 se pot vedea niște exemple, predicții generate de acest model pentru diferite gospodării. Culoarea albastră reprezintă valorile reale iar culoarea portocalie reprezintă valorile prezise. Așa cum se poate observa și din aceste grafice, rezultatele prezise sunt foarte aproape de realitate, de fapt tendința de consum de apă este urmărită (forma graficului prezis și forma graficului real este similară) doar valorile nu sunt perfect identice, dar de cele mai multe ori sunt foarte apropiate.

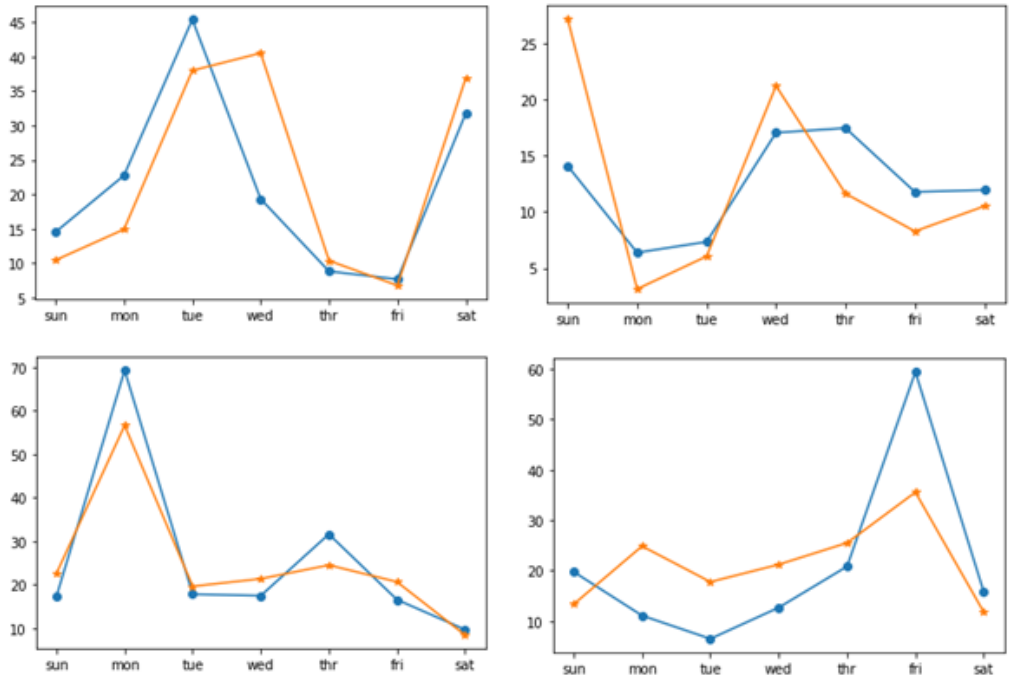


Fig. 22 Rezultatele aplicării modelului LSTM pe setul de date Helix (Chatzigeorgakidis, 2018)

4.2.3 Identificarea tiparelor de consum săptămânal (Scenariul UPB.A1)

Un alt scop al analizei datelor este de a împărți consumatorii în grupuri similare în funcție de consumul de resurse de apă. Drept urmare, metoda de clustering K-Means (Arsene et al., 2021) a fost aplicată pe setul de date Helix (Chatzigeorgakidis, 2018) cu scopul de a împărți consumatorii în grupuri similare în funcție de consumul de resurse de apă, în timp ce diferite metode de pre-procesare au fost utilizate pentru a crește nivelul de detaliu în ceea ce privește comportamentele identificate ale acestora. Acest algoritm face ca seriile de timp să fie partiționate într-un număr de clustere bazate pe algoritmul de clustering K-Means, care atribuie iterativ fiecare observație celui mai apropiat centroid în funcție de distanța euclidiană și ulterior recalculează centroizii la fiecare pas, până când un criteriu de oprire determină convergența.

Pentru a identifica numărul optim de clustere pentru setul de date disponibil, am aplicat metoda Elbow pe baza unui criteriu cunoscut sub numele de inerție sau sumă de pătrate în interiorul grupului (WCSS), care nu a dezvăluit un rezultat clar, având o curbă destul de netedă, așa cum se arată în Fig. 23.

A doua metodă de a identifica numărul optim de clustere a implicat utilizarea măsurii distorsiunii în loc de WCSS, care a dezvăluit un număr optim de 4 clustere, așa cum se arată în Fig. 24. Prin urmare, am luat în considerare 4 clustere pentru identificarea modelelor de consum din setul de date. Rezultatele inițiale sunt prezentate în Fig. 25, prezentând modele distincte în ceea ce privește forma și volumul real.

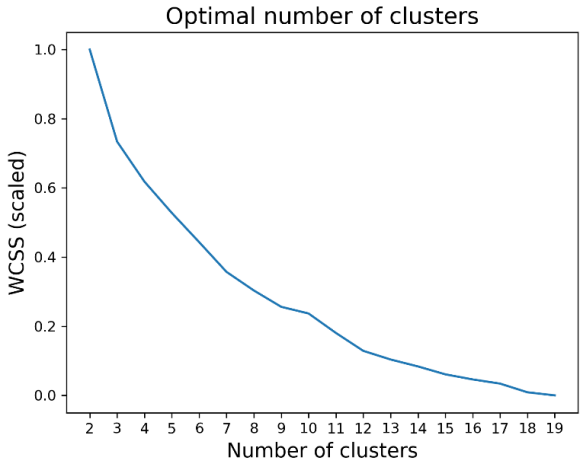


Fig. 23 Metoda Elbow pe baza WCSS

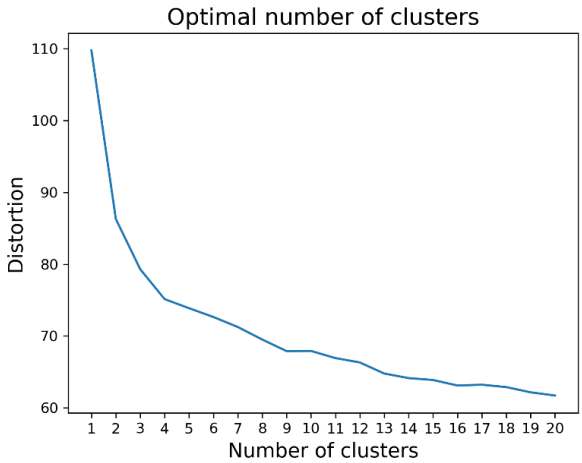


Fig. 24 Metoda Elbow pe baza distorsiunii

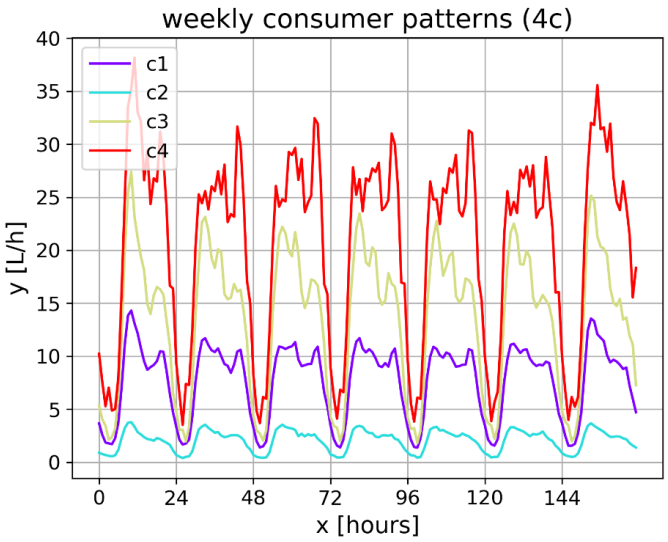


Fig. 25 Tiparele de consumatori pe baza celor 4 clustere

Deși rezultatele obținute sunt interesante, ar fi necesară o analiză mai aprofundată pentru a sublinia comportamentele consumatorilor din diferite perspective, cum ar fi tiparele efective sau tendințele și variațiile din setul de date. Pentru acest lucru a fost realizata o altă abordare în care se încearcă o identificare mai precisă a modelelor consumatorilor, folosind o etapă suplimentară de pre-procesare care implică normalizarea datelor, așa cum este prezentat în Fig. 26 și Fig. 27.

În Fig. 26, modelele săptămânale ale consumatorilor sunt prezentate pentru 4 clustere și o normalizare bazată pe sumă pentru fiecare set de date de consum. Rezultatele subliniază cantitatea de variație pe perioada săptămânală pentru fiecare tipar de consum (c1 - c4), cu c1 și c3 care prezintă variații mai mari, cu o mulțime de vârfuri, în timp ce c2 și c4 sunt mai constante în ceea ce privește consumul.

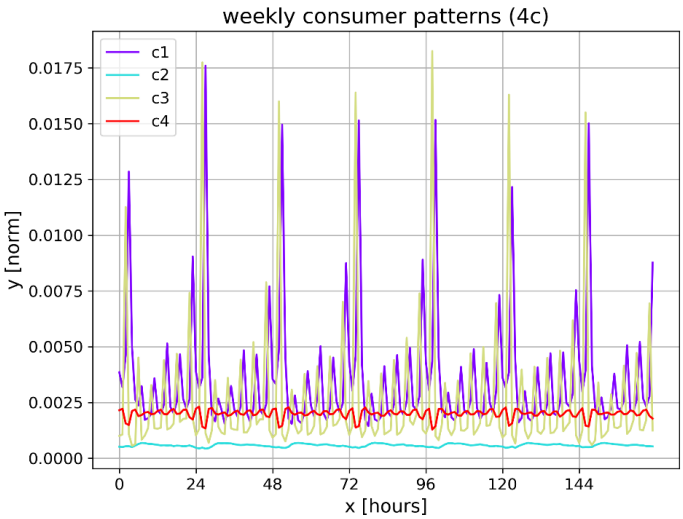


Fig. 26 Tipare de date normalizate

În Fig. 27, datele consumatorilor au fost din nou normalizate în funcție de valorile minime și maxime pentru fiecare tipar de consumator, pentru a oferi o analiză comparativă în ceea ce privește formele reale pentru fiecare tipar de consumator.

Rezultatele relevă diferențe semnificative între tipurile de consumatori identificați, atunci când se ia în considerare ora din zi după cum urmează:

- c3 arată o cerere destul de constantă în timpul zilei, de dimineața până la miezul nopții;
- c2 prezintă un model regulat cu vârfuri în jurul miezului nopții și dimineții;
- c4 arată un tipar de consum în creștere în timpul zilei cu vârfuri în timpul nopții și dimineața;
- c1 arată un tip de consum în scădere în timpul zilei, care atinge vârfurile în timpul orelor de dimineață.

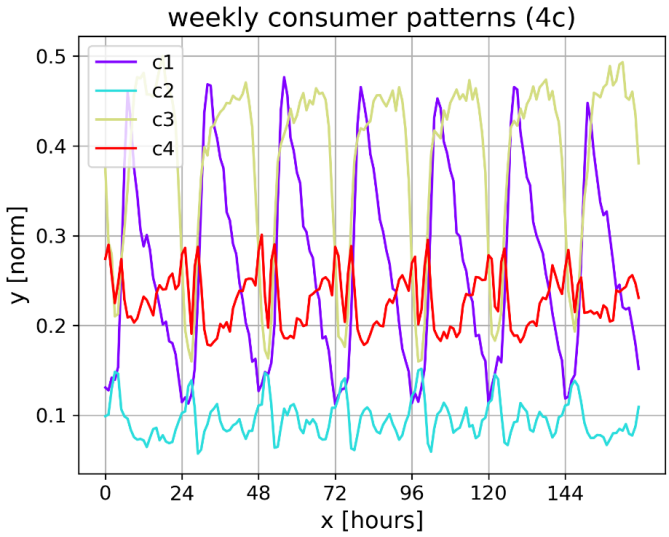


Fig. 27 Tipare din datele normalizate de 2 ori

Prin urmare, am aplicat și metoda de clustering K-Means folosind 4 clustere pentru setul de date pre-procesat. Rezultatele sunt prezentate în Fig. 28, evidențiind diferențe semnificative între tipurile de consumatori identificați atunci când se ia în considerare ziua săptămânii după cum urmează:

- c1 arată o cerere mai mare în timpul săptămânii și o cerere mai mică în weekend;
- c2 arată o cerere mai mare în weekend și o cerere mai mică în timpul săptămânii;
- c3 arată o cerere mai mare în weekend și o cerere semnificativ mai mică în timpul săptămânii;
- c4 arată o cerere constantă în timpul săptămânii și o cerere semnificativ mai mică în weekend.

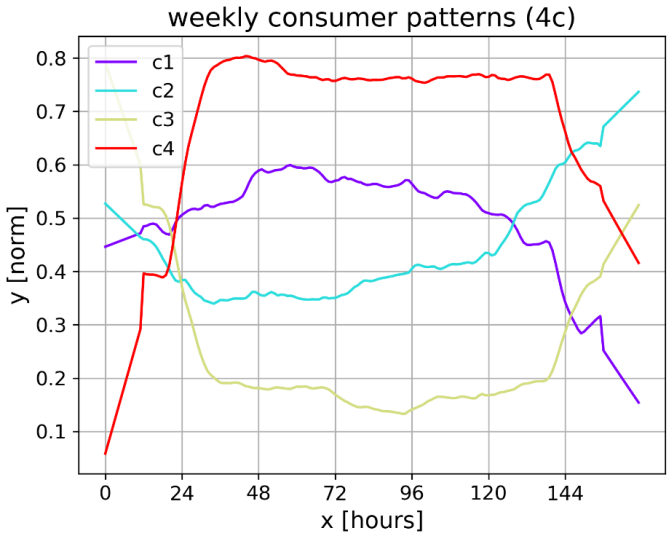


Fig. 28 Tiparele de consum normalizate dupa trendurile saptamanale

Rezultatele arată perspective diferite asupra modelelor cererii consumatorilor, care pot oferi detalii suplimentare în ceea ce privește consumul (volum, model, variație). Deoarece studiul de caz a implicat un set de date cu consumul mediu săptămânal, rezultatele relevă modele săptămânale medii care pot fi utilizate pentru a identifica comportamentul consumatorului pe întregul interval de timp.

4.2.4 Clasificarea consumatorilor după nivelul deviației standard și al geo-localizării (Scenariul UPB.A2)

Datorită faptului că setul de date Helix (Chatzigeorgakidis, 2018) conține și informații despre locația geografică a fiecărei gospodării (longitudinea și latitudinea), am găsit necesar să studiem mai în detaliu și o altă abordare prin identificarea diferitelor tipare de consumatori în funcție de locația geospațială a gospodăriilor. Pentru a atinge acest scop s-au extras clusterele din punct de vedere al regiunilor geografice în care se aflau consumatorii folosindu-se clusterizarea OPTICS. Rezultatele au arătat formarea a 7 zone geografice. Numărul optim de tipuri de consumator pentru fiecare regiune geografică a fost evaluat utilizând o implementare automată a metodei Elbow. Numărul optim de clustere pentru fiecare din zonele identificate a fost găsit între 3 și 5, așa cum se arată în Fig. 29. Metoda a dat rezultate clare pentru fiecare zonă în ceea ce privește evaluarea vizuală. Un exemplu poate fi observat în Fig. 30 pentru zona 7, având un număr optim de 3 clustere evidențiate de punctul Elbow.

Pentru extragerea comportamentului consumatorului pentru fiecare zonă, a fost utilizat algoritmul de clustering K-Means, cu numărul optim de clustere identificat prin metoda Elbow. Rezultatele evidențiază diferențele de comportament ale consumatorilor, care sunt deosebit de semnificative din punct de vedere al variației, după cum se poate observa în Fig. 31.

Fig. 29 Numărul optim de clustere pe zone

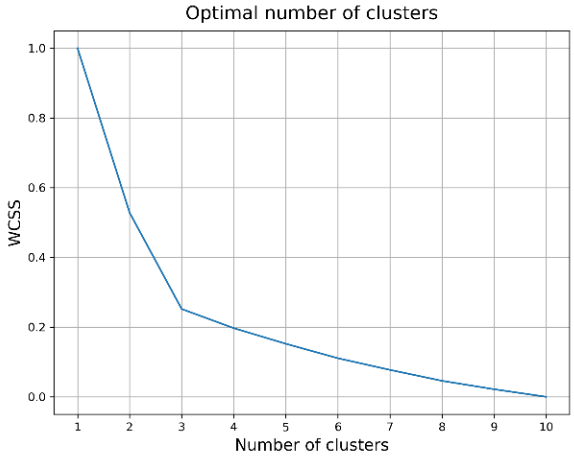


Fig. 30 Metoda Elbow aplicată pe zona 7

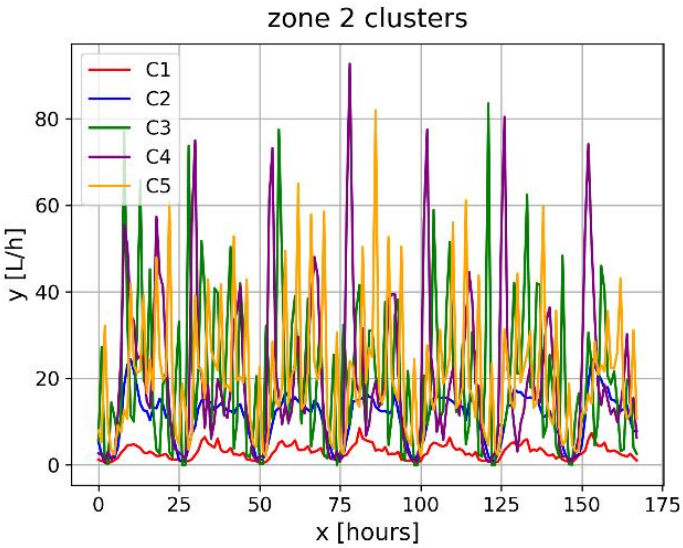


Fig. 31 Rezultatele clusterizării pe o singura zonă

În acest sens, modelele consumatorilor pentru a doua zonă de clusterizare poate fi interpretată după cum urmează:

- c1 - tipar regulat, în general consecvent peste săptămână;
- c4 - tipar regulat, cu vârfuri semnificative în timpul dimineții și seara;
- c2 - model regulat, cu volum constant în timpul zi;
- c3 și c5 - model zgomotos și neregulat, cu o mulțime de vârfuri.

Prin urmare, rezultatele au fost normalizate și s-a efectuat o analiză a seriei temporale pentru a extrage componenta sezonieră, așa cum se arată în Fig. 32, prin extragerea abaterii standard pentru fiecare zonă și model de consumator. Abaterile standard pentru componentele sezoniere au fost folosite pentru a clasifica comportamentele consumatorilor în 4 tipuri diferite, de la cea mai

consistentă, adică cea mai mică deviație standard, la cea mai variabilă, adică cea mai mare abatere standard.

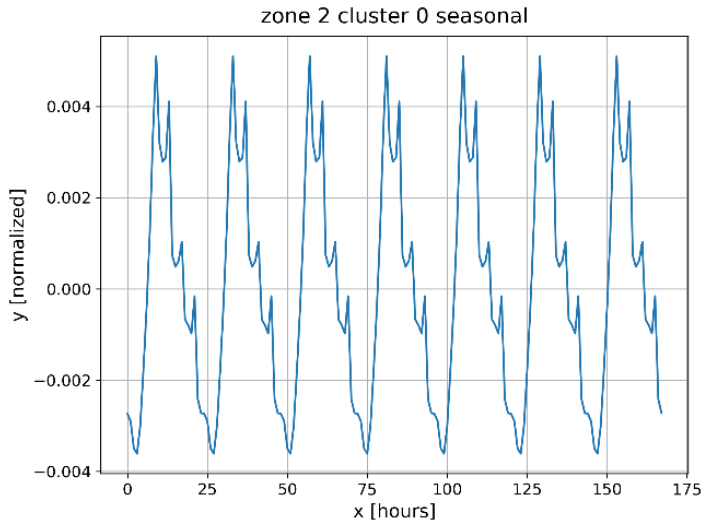


Fig. 32 Descompunerea sezonieră a datelor normalizate

Rezultatele clusterizării K-Means au fost asociate cu coordonatele geografice din setul de date original, în funcție de atribuțiile de cluster din zonele de cluster identificate în etapa de clusterizare OPTICS.

Pentru a oferi suport decizional geo-referențiat în ceea ce privește recomandările consumatorilor, rezultatele finale ale grupării sunt prezentate în Fig. 33. Vizualizarea pe hartă prezintă nivelurile de deviație standard prin dimensiunea punctelor și culoarea asociată pentru tipul de consumator. Coordonatele asociate sunt obținute din setul de date original. Rezultatele sunt corelate cu clusterele identificate din fiecare zonă, evidențiind comportamentele consumatorilor pentru întreaga rețea utilizând aceeași măsură a deviației standard.

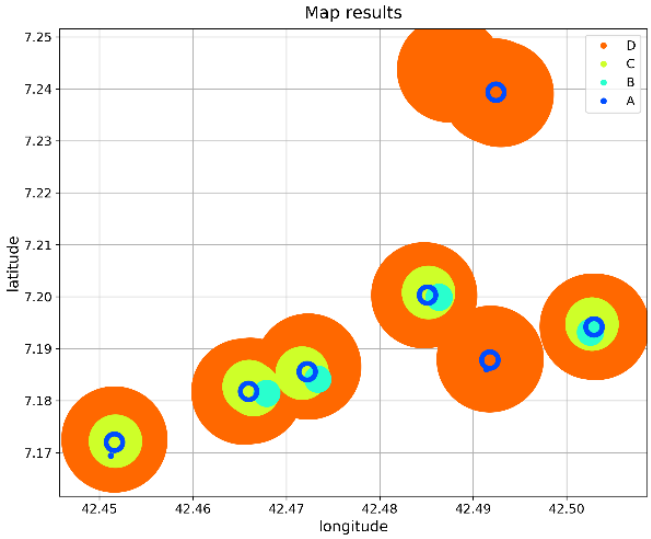


Fig. 33 Nivelul abaterii standard la nivel de cluster

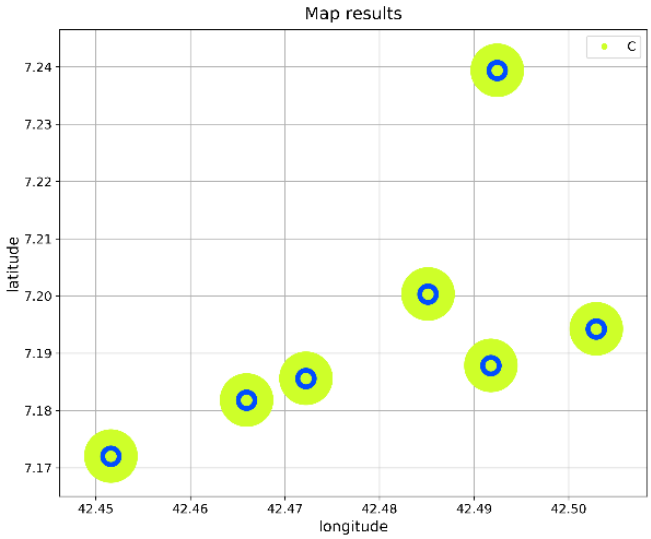


Fig. 34 Nivelul abaterii standard la nivel de zone

O altă perspectivă este prezentată în Fig. 34, în care fiecare zonă este caracterizată prin deviația standard medie extrasă din grupurile identificate. Se constată că nivelurile de deviație standard sunt similare pentru fiecare zonă, prezentând un comportament mediu al consumatorului reprezentat de tipare în mare parte zgomotoase.

În ceea ce privește tiparele de consum medii și în funcție de scara zonelor geografice, cu intervalele luate în considerare pentru deviația standard, comportamentele combinate ale consumatorilor au ca rezultat o funcționare mai puțin optimă pentru rețeaua principală de distribuție în ceea ce privește durata de viață a infrastructurii, în timp ce variațiile la nivel de consumator pot influența durabilitatea la o scară mai mică, adică rețelele rezidențiale.

Până în prezent, aceleași niveluri de deviație standard au fost utilizate pentru a caracteriza comportamentul consumatorului în întreaga rețea. Cu toate acestea, pentru a oferi rezultate mai precise în ceea ce privește comportamentul consumatorului independent, nivelurile au fost calculate pentru fiecare zonă geografică.

În mod similar, rezultatele geo-referențiate pentru suport decizional sunt prezentate în Fig. 35. În acest caz, rezultatele pot oferi o reprezentare mai precisă în ceea ce privește distribuția locală a clusterelor de consumatori pentru fiecare zonă geografică, pe baza parametrilor specifici și a condițiilor de funcționare.

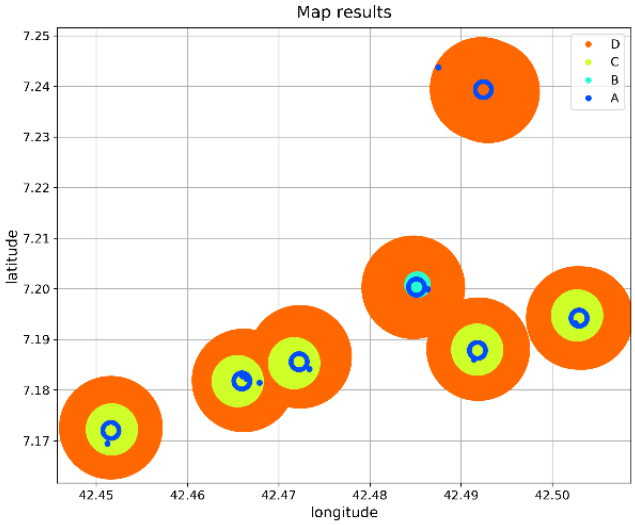


Fig. 35 Nivele standard de deviație în funcție de clustere si zone

4.2.5 Clasificarea consumatorilor după categorii predefinite (Scenariul UPB.A3)

Un alt scenariu propus se bazează pe setul de date publicat în (Austin Water, 2021). Pentru a lucra mai ușor, s-a efectuat o pre-procesare a datelor, adăugând un identificator unic pentru fiecare categorie de consumator având ca atribut codul poștal. Astfel, au fost identificate 160 de ID-uri diferite reprezentând numărul total de locuințe de unde au fost colectate datele.

Pentru a oferi o comparație paralelă între rezultatele grupării și clasele de consumatori definite în setul de date, au fost utilizate 4 clustere pentru gruparea K-Means, care au evidențiat modele distinctive în ceea ce privește variabilitatea și volumul total, așa cum se arată în Fig. 36.

Pentru fiecare categorie de consumator definită în setul de date, consumul mediu este reprezentat în Fig. 37, arătând o suprapunere între două dintre clase.

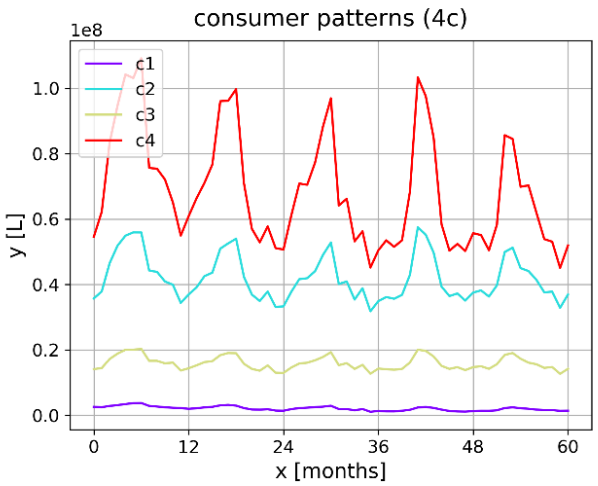


Fig. 36 Tipare de consum obținute prin clusterizare

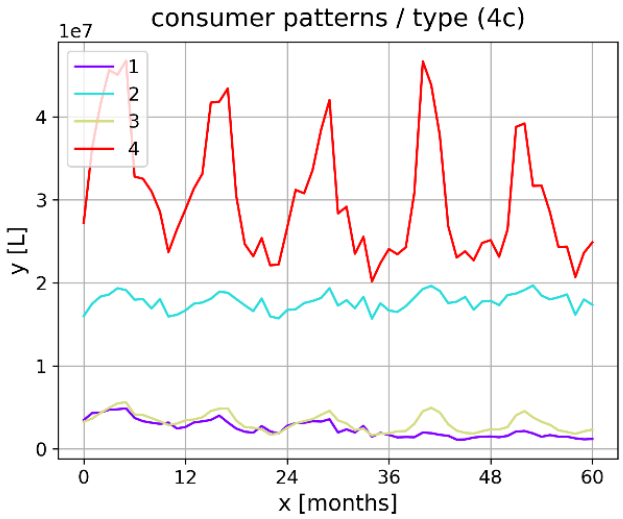


Fig. 37 Tipare de consum extrase din categoriile predefinite de consumatori

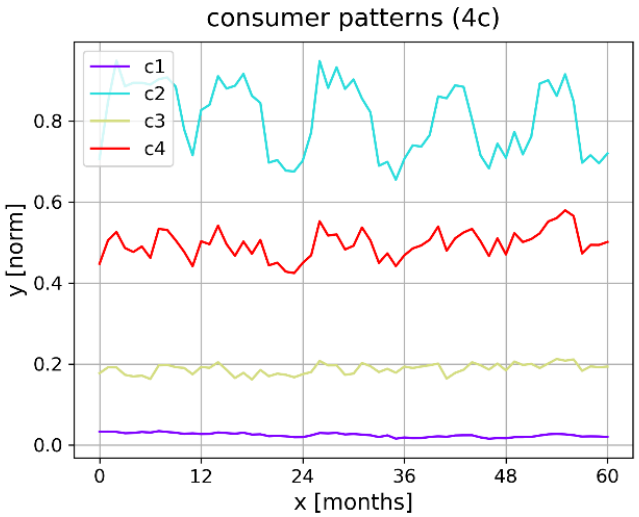


Fig. 38 Tipare de consum obținute prin clusterizare cu normalizare

Pentru a analiza în continuare posibilitatea de a găsi o similaritate și pentru a oferi o imagine de ansamblu mai bună a modelelor relative, normalizarea datelor a fost utilizată înainte de algoritmul de grupare, care a dezvăluit modele distinctive în Fig. 38.

În momentul in care se iau în considerare clasele originale, chiar și cu normalizarea datelor, c1 se caracterizează printr-un consum foarte scăzut în comparație cu celelalte tipuri, așa cum se arată în Fig. 39. Tiparele arată o imagine de ansamblu asupra comportamentului consumatorului în ceea ce privește variabilitatea consumului pe parcursul anului și volum relativ.

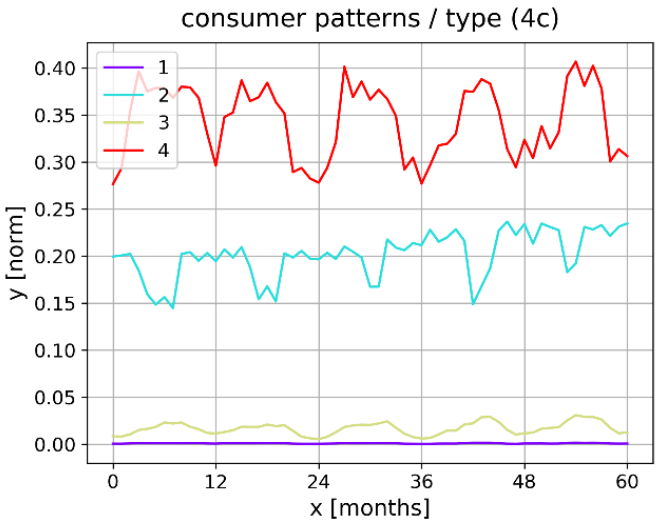


Fig. 39 Tipare de consum extrase din categoriile predefinite de consumatori cu normalizare

Se constată că normalizarea datelor nu modifică rezultatele, întărind remarcile inițiale despre asemănarea dintre sarcinile de grupare și clasele de clienți.

Mai mult, cu ajutorul algoritmilor de învățare supervizată, s-au creat modele de antrenament pentru a prezice categoria căreia îi aparține un consumator, în funcție de consumul lunar de apă. Având în vedere metoda de clasificare, am implementat și testat patru tipuri diferite de algoritmi de învățare supervizată, cum ar fi:

- Decision Tree;
- Random Forest;
- Dense Random Neural Network;
- Recurrent Neural Network (RNN).

La testarea algoritmilor s-a luat in considerare un volum de 50% din setul de date pentru date de antrenament, iar restul va fi prezis. Arborii de decizie sunt o soluție eficientă și rapidă pentru mai multe seturi de date, care sunt liniare. Mai mult, pentru a crește acuratețea, s-a testat si algoritmul Random Forest. Acesta este un proces riguros, deoarece combină mai mulți arbori de decizie, dar este mai precis, oferind un rezultat mai bun.

În continuare au fost evaluate metodele de clasificare descrise mai sus, utilizând datele țintă reprezentate de clasele de clienți. Pentru modelele Decision Tree și Random Forest, valorile țintă au fost mapate la o reprezentare întreagă, deoarece acești algoritmi pot funcționa direct cu date categorice. Pentru celelalte două modele, a trebuit să fie utilizată codificarea one-hot, cu o funcție de pierdere adecvată pentru clasificarea multi-clasă. Precizia de clasificare pentru fiecare model este reprezentată în Fig. 40, arătând o acuratețe mai mare pentru primele două modele, și o acuratețe considerabil mai mică pentru ultimele două modele.

Se observă că Decision Trees se potrivește cel mai bine pe acest set de date, cu o acuratețe puțin mai bună decât modelul Random Forest. Pentru Dense Random Neural Network și RNN, rezultatele sunt slabe, ceea ce poate fi explicat prin cantitatea insuficientă de date disponibile pentru

antrenarea modelelor. Modelul RNN este încă capabil să învețe din date mult mai bine decât modelul Dense.

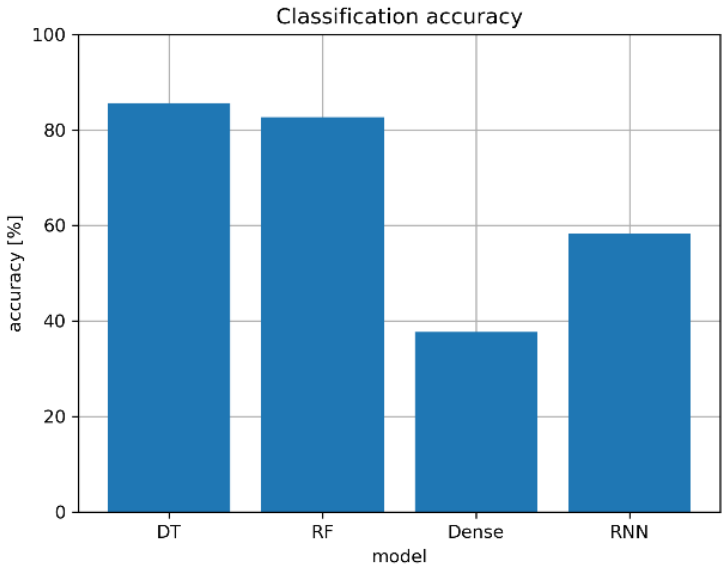


Fig. 40 Acuratețea clasificării

Rezultatele ar trebui să ofere o imagine de ansamblu asupra potențialului de grupare și clasificare într-un set de date privind consumul de apă cu clase de consumatori etichetate. Atât metodele de învățare supervizată, cât și cele nesupervizate au unele limitări, care pot fi depășite prin perspectivele complementare pe care le oferă.

Fiecare scenariu propus oferă un nivel de detaliu care poate fi potrivit pentru un anumit tip de părți interesate implicate în monitorizarea și optimizarea resurselor de apă.

5 Proiectarea componentelor și interacțiunilor

Componentele sistemului sunt proiectate separat și sunt definite prin interacțiuni specifice din perspectiva cazurilor de utilizare. Definirea interacțiunilor la nivel de sistem CPSS urmează a fi tratată în faza de integrare.

5.1 Monitorizare consum (WMS)

Componenta de monitorizare a consumului are la bază un sistem IoT de colectare a datelor de la consumatori, pe baza unor module programabile disponibile pe piață. În proiectarea modulelor de achiziție de date, am avut în vedere simplitatea soluției pentru o instalare rapidă și costuri minime.

Etapele proiectării au fost centrate pe următoarele obiective:

- Realizarea schemei electrice a modulului IoT pentru achiziția datelor;
- Proiectarea modului de conectare a modulelor IoT;
- Proiectarea arhitecturii Edge de pre-procesare a datelor;
- Proiectarea arhitecturii Cloud pentru colectarea datelor;
- Proiectarea aplicației de monitorizare și raportare pentru consumatori.

5.1.1 Componenta de monitorizare debit de apă (UTCN.C)

În nivelul de monitorizare este nevoie de unul sau mai multe debitmetre conectate la unul sau mai multe noduri de colectare, în funcție de specificul fiecărei locații de monitorizare. Specificul locației constă în faptul că pot exista unul sau mai multe puncte de măsurare, iar dacă există multiple puncte de măsurare atunci, în funcție de locația acestora (pe unul sau mai multe niveluri, de exemplu), poate fi necesară prezența mai multor noduri. Trebuie luată în considerare în faza de proiectare a sistemului de colectare a datelor dintr-o locație faptul că avem de a face cu debitmetre care se conectează cablat la noduri și din cauza aceasta există limitări de distanță, alte limitări fiind date de poziția fizică a locurilor unde pot fi montate debitmetrele și de unde pot fi accesate sursele de alimentare de tip AC sau DC pentru noduri.

În plus, de la nivelul nodurilor de colectare la care sunt conectate debitmetrele trebuie asigurată o poziționare care să permită comunicarea nodului cu punctul local de acces wireless.

Se folosește la un nivel mai ridicat puterea de procesare de la nivelul nodurilor dacă debitmetrele sunt conectate direct pentru obținerea datelor de consum în litri / minut sau litri / oră și a minimului de procesare legat de acestea.

Dacă nodurile la care sunt legate debitmetrele sunt degrevate de procesarea datelor, acestea sunt transmise direct sub forma măsurată spre un SBC (single board computer) aflat tot în cadrul rețelei de senzori, la marginea acesteia. Avantajele legate de utilizarea unui SBC țin de puterea de procesare mai ridicată, resursele de memorie mai mari și de posibilitatea de stocare locală a unei cantități considerabile de date pentru crearea și menținerea unui istoric de consum.

Partea de achiziție a datelor prin intermediul nodurilor la care sunt conectate debitmetrele rămâne neschimbată față de cazul anterior, iar aici la marginea rețelei se utilizează un dispozitiv embedded care utilizează un SoC (System on Chip) și integrează în componenta centrală de procesare o arhitectură de tip ARM.

Un astfel de sistem are cerințe de consum energetic foarte mici și este capabil să accelereze aplicațiile de ML.

5.1.2 Proiectarea modulelor de achiziție (UPB.C)

Schema electrică a modulului de achiziție a datelor este prezentată în Fig. 41. Conectarea debitmetrelor YF-S201[12] se realizează prin:

- cablul de date (galben), conectat la un pin digital de pe modulul NodeMCU;
- cablul de alimentare (roșu), conectat la pinul Vin de pe modulul NodeMCU;
- cablul de masă (negru), conectat la pinul GND de pe modulul NodeMCU.

Dintre acestea, cablurile de alimentare (și masă) sunt unite pe un singur pin la intrarea modulului, fapt ce necesită lipirea manuală a conectorilor terminali și presupune o mare parte din efortul necesar pentru instalare.

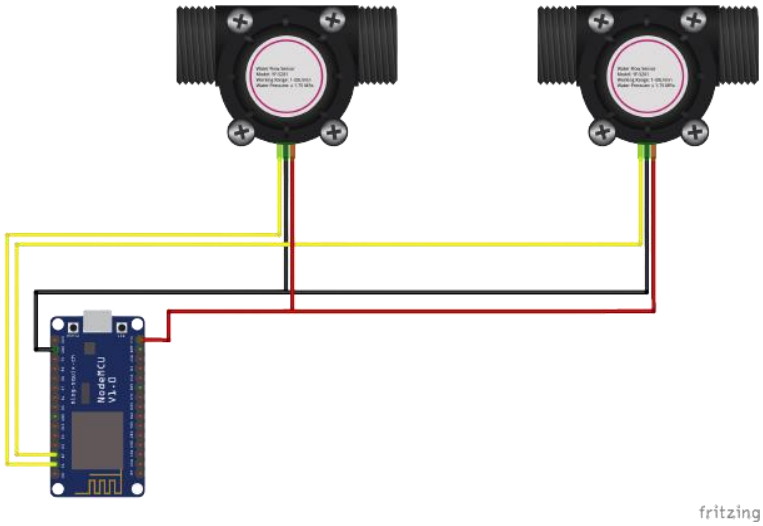


Fig. 41 Schema electrică a modulului de achiziție a datelor

Modulele NodeMCU sunt alimentate prin cablu USB de la un alimentator de rețea. Deoarece acestea sunt proiectate să transmită date cu o frecvență ridicată, nu este posibilă funcționarea pe baterie. Au fost achiziționate cabluri USB lungi de 5 m pentru a permite alimentarea modulelor în locuințe, unde prizele de alimentare pot fi la distanțe mai mari de locul instalării senzorilor.

Procesul de instalare constă în pre-programarea modulelor și configurarea acestora la locul instalării, pentru a se conecta la rețeaua locală de WiFi și a transmite conform cu senzorii conectați la pinii digitali.

5.1.3 Componenta Edge de preprocesare a datelor (UTCN.C)

Pe nivelul edge, datele trec printr-o etapă de screening prin care sunt eliminate toate valorile invalide apărute din cauza calibrării debitmetrelor, a degradării capacității de măsurare cu acuratețe sau datorate întreruperilor în alimentarea cu energie electrică a acestora.

[12] https://www.optimusdigital.ro/ro/senzori-senzori-hall/3078-debitmetru-yf-s201-1-30-l-min.html

După aceasta se trece la etapa de filtrare a citirilor debitmetrelor de tip outlier cauzate de factori incontrolabili, cum sunt impuritățile de la nivelul fluxurilor de apă măsurate, care influențează funcționarea componentelor în rotație din cadrul debitmetrelor.

Pentru debitmetrele comerciale disponibile pentru măsurarea consumului rezidențial/casnic, acestea sunt capabile sa înregistreze valori ale debitului în intervalul de 1 litru – 40 litri / minut. Pentru eventualele valori pe care debitmetrele le transmit în afara acestui interval, nodurile de tip NodeMCU efectuează filtrarea valorilor considerate în afara intervalului de referință.

De asemenea, în ceea ce privește eșantioanele lipsă în cadrul fluxului de măsurare pe un debitmetru la nivelului nodului se aplică funcția de aproximare a valorilor lipsă din cadrul secvenței de valori măsurate. Pentru a realiza acest lucru este posibilă utilizarea unei funcții de interpolare care realizează o medie aritmetică a *n* valori ale eșantioanelor primite înainte de eșantionul lipsă și a *n* valori primite după eșantionul lipsă, unde *n* este un factor configurabil la nivelul fiecărui nod în parte în funcție și de valorile obținute pentru debit în cadrul etapei de calibrare.

Tot la nivelul unui nod are loc etapa de agregare a tuturor eșantioanelor primite de la debitmetrele conectate local la un nod. În cazul punctelor de consum apropiate fizic cum sunt cele din încăperile de mici dimensiuni se realizează o agregare a tuturor valorilor de consum și se realizează o singură raportare către componenta aflată la marginea edge-ului pentru o transmisie mai departe în cloud.

Tot la nivelul acestei componente se realizează în funcție de specificul aplicației și etapa de reducere a valorilor redundante.

5.1.4 Serviciul de colectare a datelor (UTCN.C)

Serviciul de colectare a datelor are două componente (Fig. 42):

- API de colectare / regăsire a datelor;
- Baza de date.

API-ul poate fi accesat de dispozitivele de monitorizare / edge și de alte aplicații care procesează datele prin cereri HTTP de tip GET și POST. Cererile trimise înspre API și răspunsurile acestuia folosesc date împachetate în format JSON.

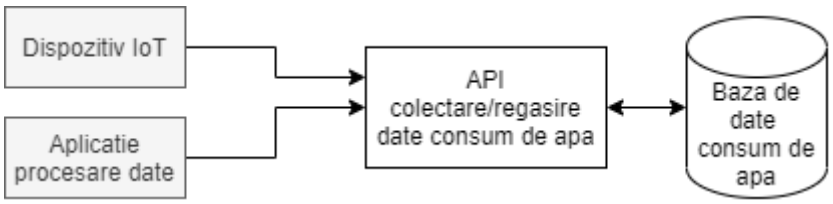


Fig. 42 Serviciul de colectare a datelor

Interacțiunea între dispozitivele de monitorizare / edge și serviciul de colectare a datelor este descrisă în Fig. 43. Dispozitivele trimit periodic datele încapsulate într-o cerere HTTP de tip POST, în format JSON. Datele preluate în acest fel sunt stocate în baza de date.

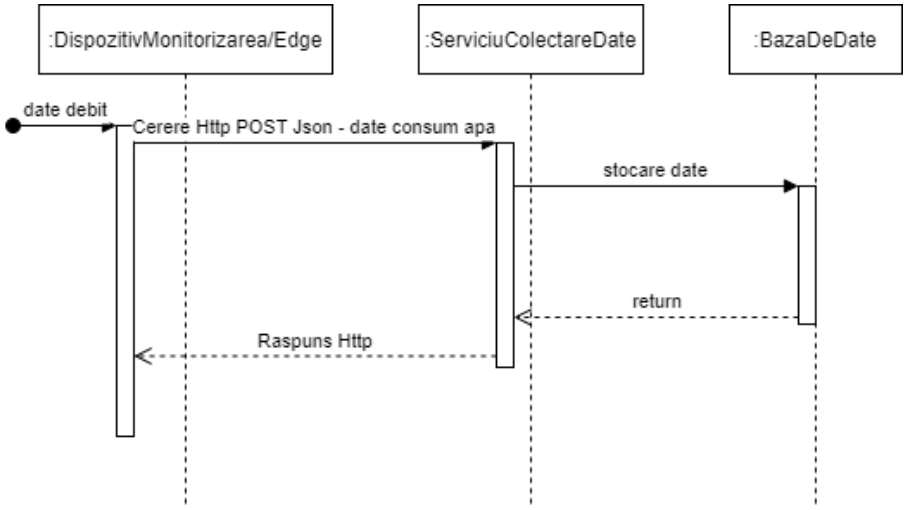


Fig. 43 Interacțiunea între un dispozitiv de monitorizare/edge și serviciul de colectare a datelor

Datele colectate pot fi accesate de aplicații client, cum ar fi, de exemplu, una care face procesarea / analiza datelor. Aplicația de procesare trimite o cerere HTTP de tip GET, după care primește datele ca răspuns, în format standard SensorThings. Interacțiunea între aplicația de procesare a datelor și serviciul de colectare a datelor este descrisă în Fig. 44.

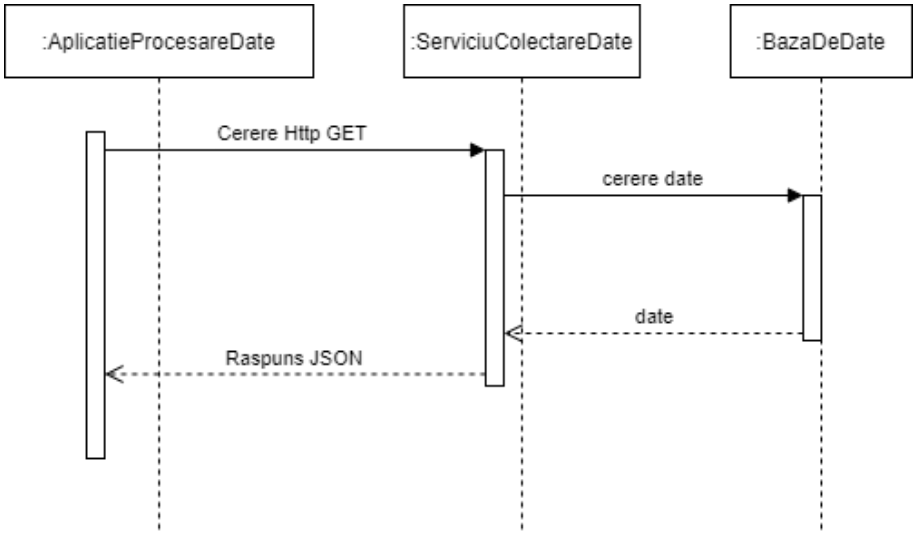


Fig. 44 Interacțiunea între o aplicație care face procesarea datelor și serviciul de colectare a datelor

Modelul folosit pentru preluarea, stocarea și servirea datelor este cel specificat în standardul OGS SensorThings și descris în Fig. 45. Datele transmise între clienți și API sunt structurate ca *datastreams*. Datele transmise vor conține atâtea datastreams câte puncte de măsurare sunt instalate la o locație. Dacă la o locație există un singur punct de măsurare al consumului de apă, datele vor fi structurate într-un datastream, ca în exemplul următor.

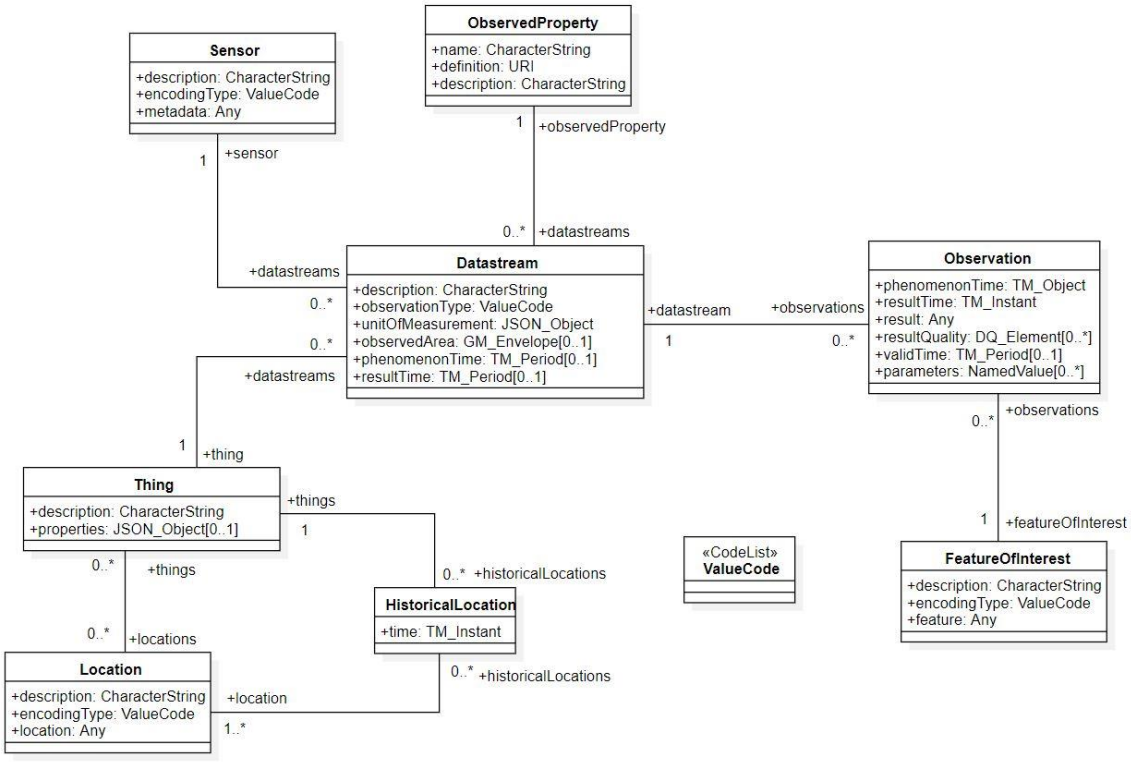


Fig. 45 Modelul de date OGC SensorThings

ExempluConsum.JSON

```
{
	"name": "Lab-test-debitmeter",
	"description": "lab-test",
	"properties": {
		"style": "lab"

	},
	"Locations": [
		{
			"name": "lab-C04",
			"description": "lab baritiu 28",
			"encodingType": "application/vnd.geo+json",
			"location": {
				"type": "Point",
				"coordinates": [23.586384084911145, 46.77302048938485]
			}
		}
	],
```

```
	"Datastreams": [
		{
			"name": "test-debitmeter",
			"description": "consum apa rece",
			"observationType": "http://www.opengis.net/def/observationType/OGC-OM/2.0/OM_Measurement",
			"unitOfMeasurement": {
				"name": "litre",
				"symbol": "L",
				"definition": "http://www.qudt.org/qudt/owl/1.0.0/unit/Instances.html#Liter"
			},
			"Sensor": {
				"name": "YF-S201",
				"description": "Debit meter-water",
				"encodingType": "application/html",
				"metadata": "https://www.optimusdigital.ro/ro/senzori-senzori-hall/3078-debitmetru-yf-s201-1-30-l-min.html"
			},
			"ObservedProperty": {
				"name": " water consumption ",
				"definition": "http://www.qudt.org/qudt/owl/1.0.0/quantity/Instances.html#Volume",
				"description": "consum de apa"
			},
			"Observations": [
				{
					"phenomenonTime": "2021-09-14T10:04:00Z",
					"result": 2
				},
				{
					"phenomenonTime": "2021-09-14T10:05:00Z",
					"result": 2.5
				}
			]
		}
	]
}
```

5.1.5 Aplicația mobilă (UPB.WMS)

Aplicația mobilă permite furnizorilor și consumatorilor să vizualizeze informații despre consumul de apă din locuință. Pentru furnizor, aplicația permite monitorizarea senzorilor instalați, adăugarea unui nou senzor și asocierea cu un consumator pe baza unui cod de client, debranșarea, precum și vizualizarea notificărilor transmise de consumatori cu privire la problemele ce pot apărea în furnizarea apei.

Consumatorul beneficiază de un set de funcționalități mai restrânse pentru monitorizarea senzorilor asociați, precum și actualizarea datelor personale și transmiterea notificărilor pentru a semnala anumite probleme.

Aplicația se conectează la server-ul aplicației pentru sincronizarea datelor de la senzorii instalați, apoi se abonează la brokerul MQTT pentru actualizarea și vizualizarea acestora în timp real.

Datele de identificare ale utilizatorilor precum și funcționalitățile ce presupun autentificarea și vizualizarea datelor de consum sunt gestionate de serviciul Firebase, separat de server-ul aplicației. Firebase Realtime Database (Google, 2021a) este o bază de date ne-relațională, având o structură arborescentă de tip JSON (JavaScript Object Notation) și o interfață de acces programat cu suport optimizat pentru tehnologii mobile (Moroney, 2017).

Date de consum

Pentru vizualizarea datelor istorice, se realizează o cerere HTTP iar server-ul aplicației răspunde cu datele înregistrate în baza de date MySQL folosind un set de filtre (e.g. ID-ul modulului, canalul de măsurare, numărul de înregistrări returnate sau intervalul de timp solicitat). În mod normal datele sunt actualizate în baza de date la intervale egale de timp, ceea ce permite afișarea lor pe grafic, fără o prelucrare intermediară.

Notificări

Prin funcția de notificare, furnizorul poate primi de la consumatori notificări despre problemele raportate de aceștia. Acestea pot fi:

- reclamații;
- solicitări intervenții;
- instalare de senzori.

Prin reclamații, consumatorul poate semnaliza anumite probleme în furnizarea apei precum: oprirea alimentării cu apă, probleme de facturare, branșamente ilegale, etc.

Prin solicitări de intervenție, consumatorul poate solicita intervenția unei echipe autorizate pentru operații de mentenanță sau reparații în sistemul de distribuție.

Notificarea de tip instalare de senzori este specifică aplicației de monitorizare a consumului, prin care un nou client poate solicita configurarea modulelor instalate, fiind transmise date de identificare ale solicitantului în vederea actualizării datelor din aplicație.

5.2 Suport decizional (DSS)

În contextul DSS propus, scenariile de evaluare sunt independente, iar metodele de analiză propuse pot fi utilizate la cerere de către părțile interesate. Având în vedere rolurile multiple și părțile interesate într-o rețea de distribuție a apei, soluția propusă poate oferi suport decizional și recomandări după cum urmează:

- Atât consumatorii, cât și furnizorii și operatorii de rețea pot primi alerte în momentul în care se detectează anomalii în datele culese de serviciul de colectare a datelor, alerte pe baza cărora să se poată lua diverse decizii (de ex. înlocuirea unor echipamente sau investigarea integrității implementării).
- Furnizorii pot primii estimări ale consumului ce urmează a fi generat de diverși consumatori în parte și să ia decizii pe baza acestor estimări (de ex. debranșarea în cazul în care datoriile ar depăși anumite limite maxime).
- Furnizorii și operatorii de rețea pot beneficia de o mai bună înțelegere a comportamentului consumatorului în scopul dezvoltării de soluții stimulative și centrate pe consumatori pentru optimizarea resurselor de apă.
- Furnizorii și operatorii de rețea pot beneficia de suport în procesul de luare a deciziilor, analizând datele extrase și profilurile consumatorilor.
- Consumatorii pot beneficia de monitorizarea propriului consum, învățând să se autoeduce și să utilizeze în mod responsabil resurselor de apă.
- Consumatorii pot beneficia de recomandări specifice în ceea ce privește zona lor geografică, prin comparație cu vecinii lor.
- Furnizorii pot beneficia de recomandări specifice cu privire la întreaga rețea, care arată o comparație cu clustere similare de consumatori din orice zonă geografică.
- Furnizorii și operatorii de rețea pot beneficia de recomandări generale pentru zonele geografice, subliniind efectul general asupra rețelei de distribuție a apei.

În plus, DSS poate oferi recomandări bazate pe nivelurile de deviație standard ale comportamentului utilizatorilor după cum urmează:

- recomandări generale pentru zonele geografice;
- liniile directoare specifice pentru grupurile de consumatori cu privire la întreaga rețea;
- recomandări specifice pentru grupurile de consumatori în ceea ce privește zona geografică a acestora.

5.2.1 Serviciile de detecție a anomaliilor (UTCN.A1) și de predicție (UTCN.A2)

Serviciile de detecție a anomaliilor și de predicție (UTCN.A1 și UTCN.A2) sunt oferite prin intermediul platformei numite AutomaticAI (Czako, Sebestyen & Hangan, 2021). Dezvoltarea acestei platforme a fost realizată în cadrul unui proiect anterior și cuprinde toate metodele și algoritmii necesari pentru a configura și regla un algoritm de detectare a anomaliilor de la zero. Platforma conține algoritmi de preprocesare, cum ar fi diferite tipuri de transformare a datelor sau metode de scalare a datelor, algoritmi de selectare a caracteristicilor și de reducere a dimensionalității, cum ar fi PCA, LDA și alții, algoritmi de clasificare și regresie și, de asemenea, o mare varietate de metode

de vizualizare a datelor și a rezultatelor. Arhitectura de nivel înalt al acestei platforme se regăsește în Fig. 46.

Pentru procesarea și analiza datelor de consum de apă și pentru detecția anomaliilor în calitatea apei s-a extins platforma existentă prin adăugarea unei funcționalități extrem de utile legate de posibilitatea de a crea și configura diferite pipeline-uri de procesare și analiză a datelor de intrare. Folosind această opțiune a fost creată soluția pentru detecția anomaliilor în calitatea apei, construind un pipeline care conține mai mulți pași de preprocesare și un model de AI pentru clasificare.

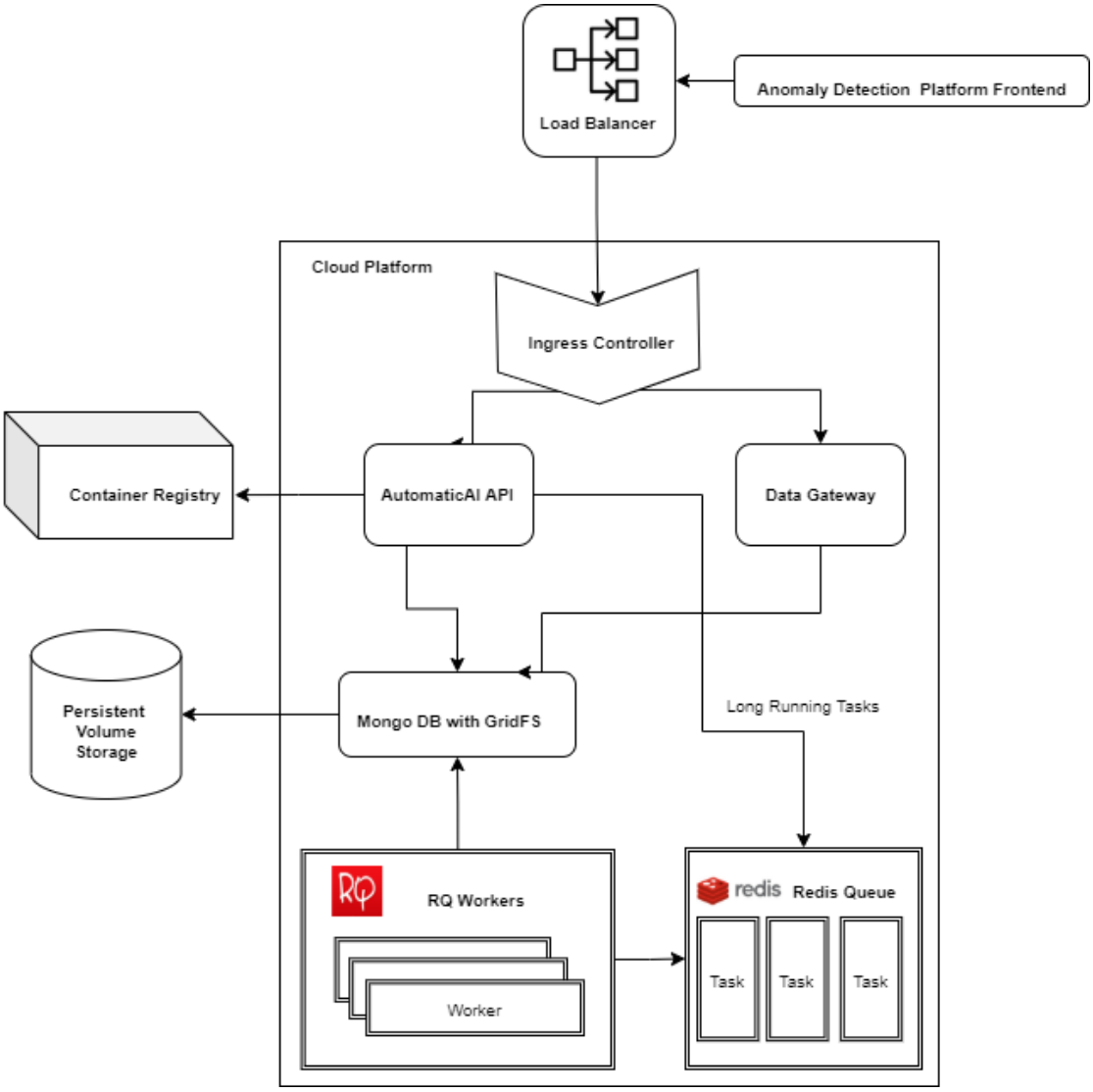


Fig. 46 Arhitectura de nivel înalt a platformei AutomaticAI

Pe lângă acest pipeline dinamic, pentru a avea funcționalități de analiză și procesare a seriilor de timp, au fost adăugate module noi pentru vizualizarea datelor de tip serii de timp, pentru configurarea modelelor de predicție, vizualizarea rezultatelor și evaluarea modelelor antrenate. Pentru predicție a fost adăugat algoritmul Auto-ARIMA (Yermal & Balasubramanian, 2017) și posibilitatea de a configura diferite arhitecturi de rețele neuronale recurente de tip LSTM (Lu et al., 2020), (Xingjian et al., 2015).

Toate aceste module și funcționalități noi au fost adăugate în serviciul AutomaticAI și a fost creată și o parte de interfață cu utilizatorul (UI) pentru folosirea cât mai ușoară a acestora.

5.2.2 Clustering pe bază de date de consum folosind K-Means (UPB.A1)

Metoda de clustering K-Means a fost utilizată pe setul de date Helix (Chatzigeorgakidis, 2018) pentru a împărți consumatorii în grupuri similare în funcție de consumul de resurse de apă, în timp ce diferite metode de pre-procesare au fost utilizate pentru a crește nivelul de detaliu în ceea ce privește comportamentele identificate ale acestora. Pipeline-ul de clusterizare este prezentat în Fig. 47.

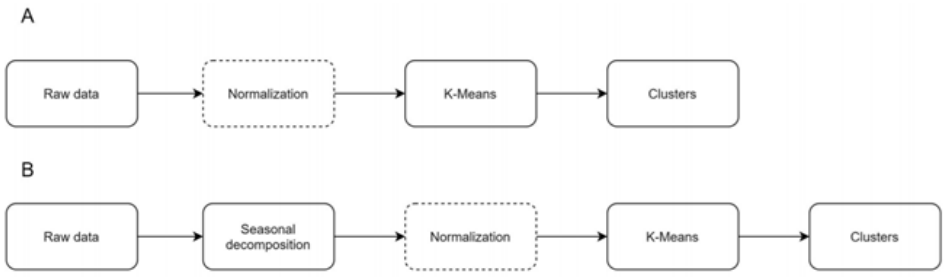


Fig. 47 Pipeline clusterizare

5.2.3 Clustering pe bază de coordonate geografice folosind OPTICS (UPB.A2)

Pentru a clasifica tipurile de consumatori pentru diferite zone geografice, identificate prin geo-locația acestora s-a utilizat clusterizarea OPTICS pornind de la setul de date Helix (Chatzigeorgakidis, 2018), rezultatele evidențiind 7 clustere bazate pe coordonatele geografice.

OPTICS (Ordering Points To Identify the Clustering Structure) are multe asemănări cu algoritmul DBSCAN și poate fi considerat o generalizare a DBSCAN, cu diferența că păstrează ierarhia clusterului pentru o rază variabilă de vecinătate. Diferența cheie dintre DBSCAN și OPTICS este faptul că algoritmul OPTICS construiește un graf de accesibilitate, care atribuie fiecărui eșantion atât o distanță de accesibilitate, cât și un punct în cadrul atributului ordering. În plus, OPTICS este mai potrivit pentru utilizarea pe seturi de date mari decât implementarea DBSCAN.

Pentru fiecare grup, s-a efectuat o analiză a seriei temporale pentru a extrage componenta sezonieră din care a fost calculată deviația standard. DSS se bazează pe multiple strategii de clasificare care implică evaluarea nivelurilor de deviație standard în contextul unei rețele de distribuție a apei. Etapa de clasificare a implicat reatribuirea fiecărui cluster pe baza nivelurilor identificate în termeni de deviație standard, pentru a evalua tiparele de consum dintr-o perspectivă numerică.

Prin urmare, modelele de suport al deciziilor sunt definite pe baza nivelurilor de deviație standard care pot fi fixate pentru întreaga rețea sau pot varia în funcție de zonă, pentru a oferi recomandări pentru diferite tipuri de părți interesate. În cele din urmă, rezultatele au fost prezentate sub formă de hartă, subliniind tipurile de consumatori, pentru a oferi suport decizional spre obținerea unor comportamente mai sustenabile ale consumatorilor.

5.2.4 Metode de clasificare (UPB.A3)

În procesul de colectare a datelor, seturile de date pot conține și valori nule (spații libere, NaN, nedefinit etc.). Acest lucru este incompatibil cu estimatorii scikit-learn, deoarece toate valorile matricei ar trebui să fie numerice, fără spații. Una dintre soluțiile la acest inconvenient este imputarea valorilor lipsă, cu alte cuvinte, încercând-se încearcă prezicerea lor folosind date cunoscute. Clasa SimpleImputer a fost folosită pentru a codifica valorile lipsă. Pentru a oferi o evaluare relativă în ceea ce privește modelele de consum, a fost utilizată normalizarea datelor. În procesul de normalizare, datele au fost scalate la intervale stabilite, folosind clasa MinMaxScaler.

Pipeline-ul de procesare folosit în scenariile UPB.A1 - UPB.A3 (Fig. 48) a fost implementat în Python folosind pachetul scikit-learn (Pedregosa et al., 2011; scikit-learn developers (BSD License)., n.d.) pentru metodele de clustering (e.g. K-Means și OPTICS), în timp ce pre-procesarea implică pachetul statsmodels[13] pentru analiza seriilor temporale (adică descompunerea sezonieră) și *numpy* pentru normalizarea datelor.

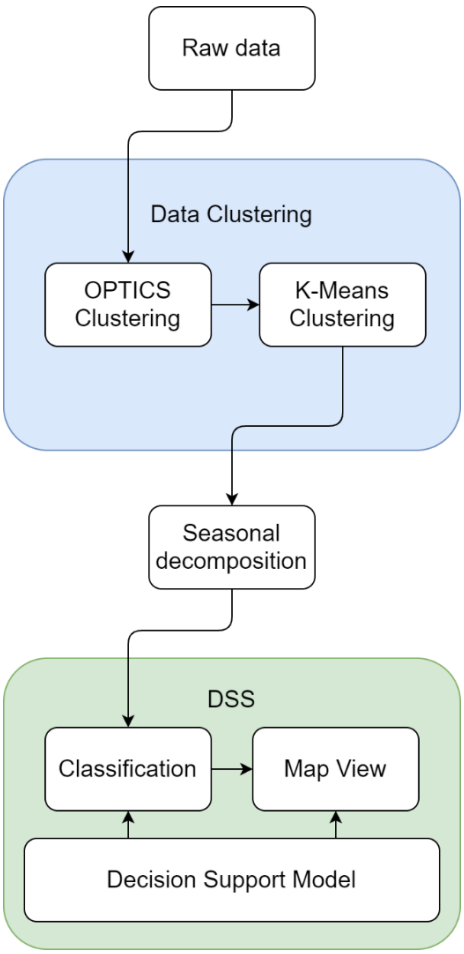


Fig. 48 Proiectarea componentei de suport decizional folosită în scenariile UPB.A1- UPB.A3

[13] https://www.statsmodels.org/stable/index.html

5.3 Crowdsensing (MCS)

În proiectarea componentei Crowdsensing, principalul obiectiv a fost de modelare a problemei într-o formă standard. Problema raportărilor în contextul MCS necesită o abordare teoretică pentru a stabili modul de alocare a sarcinilor în funcție de tipul de utilizator. În condiții reale, o astfel de aplicație presupune participarea utilizatorilor, fapt pentru care proiectarea componentei nu este doar o problemă tehnică.

Este necesar de luat în calcul modul în care sunt alocate activitățile de raportare în funcție de profilul utilizatorilor, pentru a asigura buna funcționare a soluției în ansamblu. În acest sens, pot exista probleme de ordin tehnic: scalabilitate, complexitate, tehnologii emergente, dar și non-tehnic: motivație, stimularea participării, raportări false.

Din perspectivă teoretică, o astfel de problemă se poate reduce la un scenariu de optimizare, pentru a obține un echilibru dintre priorități și reputația utilizatorilor. Alocarea activităților de raportare (identificare probleme, soluționare) poate fi modelată ca o problemă a rutării vehiculelor (VRP) adaptată la scenariul de joc, pentru o abordare generală centrată pe utilitatea platformei. Din perspectiva motivației și a costurilor individuale, un mecanism pe bază de licitații Vickrey-Clarke-Groves (VCG) reprezintă o abordare centrată pe utilizator (Hanum et al., 2018; Makowski & Ostroy, 1987).

O astfel de soluție nu poate fi evaluată decât în contextul unei aplicații pe scară largă, cu implicarea unui număr mare de utilizatori pentru a obține rezultate relevante. De aceea, pentru un model de laborator, aducerea problemei la o formă standard (VRP, VCG) permite proiectarea componentei pe baza unor soluții existente, deja validate în alte studii, și reprezintă principalul element de inovație în contextul MCS.

Problema raportărilor prin crowdsensing poate fi abordată ca o problemă de rutare în două etape. Alocarea sarcinilor este tradusă într-o problemă de rutare a vehiculelor (VRP) cu constrângeri de capacitate, în timp ce mecanismul de stimulare este perfecționat în continuare folosind un model de licitație Vickrey-Clarke-Groves (VCG) pentru raportare veridică.

5.3.1 Model de alocare a raportărilor bazat pe VRP

Modelul VRP reprezintă abordarea centrată pe platformă pentru MCS. Deși există mai multe variații ale VRP utilizate pentru rezolvarea diferitelor probleme, modelul este adaptat la scenariul MCS propus în arhitectura componentei de Crowdsensing, prin definirea contextului și a constrângerilor necesare.

Modelul de alocare a sarcinilor bazat pe VRP este definit după cum urmează: participanții sunt reprezentați de vehicule, cu capacități în funcție de nivelurile de reputație, în timp ce constrângerile de distanță definesc, de exemplu, distanța maximă de parcurs. Sarcinile sunt reprezentate de clienți, cu locații fixe și cerințe stabilite în funcție de prioritățile atribuite. Fiecare vehicul are un punct de plecare, care corespunde unui depozit, în timp ce celelalte noduri sunt asociate clienților (Hanum et al., 2018).

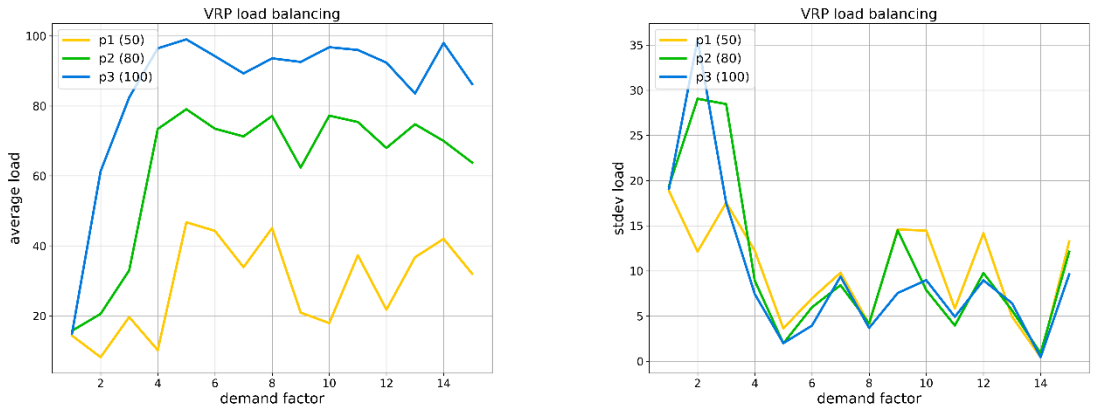
Pentru un scenariu de tip joc serios pe bază de geolocație, în care un set de provocări sunt distribuite pe hartă, proiectarea jocului pentru realizarea unei distribuții optime a participanților poate

fi definită ca o problemă de optimizare a expedierii combinată cu un mecanism de stimulare bazat pe recompense. Sarcinile implică raportarea problemelor în contextul apei urbane și pot fi extinse pentru multe cazuri de utilizare.

Cu toate acestea, pentru a obține o utilitate generală a platformei, stimulentul bazat pe recompense trebuie să fie contrabalansat de o strategie de reglementare. Scorul de reputație este moneda principală în joc și trebuie definit pentru a obține stimulente optime pentru raportări, reducând în același timp riscul de supra-raportare sau raportare falsă.

În acest sens, algoritmul VRP poate fi configurat prin constrângerile de capacitate pentru fiecare participant în funcție de reputația acestuia. La fiecare iterație a algoritmului de optimizare, capacitățile sunt actualizate. În cazul raportării veridice, capacitatea este mărită în funcție de recompensă, ceea ce permite direcționarea către sarcinile următoare. Am luat în considerare trei tipuri de participanți: potrivire 100% (raportare veridică și exactă), potrivire 80% (raportare veridică și în mare parte exactă) și potrivire 50% (raportare falsă / inexactă).

Pentru a evalua modelul și diferitele tipuri de comportament, rulăm problema de optimizare pe întreaga gamă de distribuții în ceea ce privește coordonatele geografice ale participanților. Am luat în considerare coordonatele fixe pentru sarcini și coordonatele variabile de pornire pentru participanți, pentru a evalua costul mediu al sistemului (distanța totală parcursă) în funcție de comportamentul fiecăruia și, prin urmare, pentru a evalua utilitatea platformei pentru încurajarea participării sincere.



(a) (b)

Fig. 49 Alocarea activităților prin VRP (a) Alocare medie; (b) Deviație standard.

Modelul de optimizare este implementat în Python, bazat pe framework-ul Google OR-Tools, care oferă o formulare generală a problemei VRP cu mai multe aspecte și constrângeri (Google, 2021b). Pentru a evalua echilibrarea încărcării VRP, adică alocarea sarcinilor, am luat în considerare un set de 10 locații din București și trei participanți care corespund comportamentului simulat: raportare precisă, raportare în mare parte exactă și raportare falsă. Modulul de rutare este inițializat cu matricea distanțelor, numărul de participanți și locația depozitelor (locațiile de pornire).

Pentru experimentul prezentat în Fig. 49, am evaluat efectul acestui factor de recompensă în intervalul de la 1 la 15, unde un factor de 5 are ca rezultat o distribuție globală echilibrată (recompense totale vs reputația totală). Deoarece locația de pornire a participanților poate afecta

acuratețea rezultatelor, am considerat o inițializare aleatorie folosind un cerc de delimitare. Algoritmul VRP este evaluat pentru 1000 de iterații, iar sarcina medie (alocarea resurselor) este prezentată pentru fiecare tip de participant. Conform rezultatelor, în timp ce problema de optimizare este rezolvată în orice mod, factorul optim de recompensă ar trebui definit pentru a obține rezultatul dorit pentru mecanismul de stimulare. Prin urmare, ar trebui să existe un echilibru între reputația participanților și recompensele alocate pentru fiecare sarcină disponibilă.

Prin urmare, geolocalizarea și alte constrângeri pot servi drept filtru pentru trimiterea sarcinilor către participanți pentru a realiza o strategie de rutare rentabilă pentru o utilitate sporită a platformei în ceea ce privește raportarea veridică și calitatea raportărilor.

5.3.2 Model de licitație bazat pe VCG pentru raportare veridică

Abordarea centrată pe utilizator bazată pe VCG este definită în contextul platformei MCS propuse. În licitația VCG, există mai multe articole și agenți, fiecare agent plasând o ofertă pentru fiecare articol, fără a cunoaște evaluările altora. Sistemul de licitație atribuie articolele în funcție de evaluările lor reale, în timp ce fiecare agent plătește doar valoarea marginală în comparație cu ceilalți agenți, încurajând licitarea veridică (Makowski & Ostroy, 1987; Sessa et al., 2017; Tao & Song, 2018).

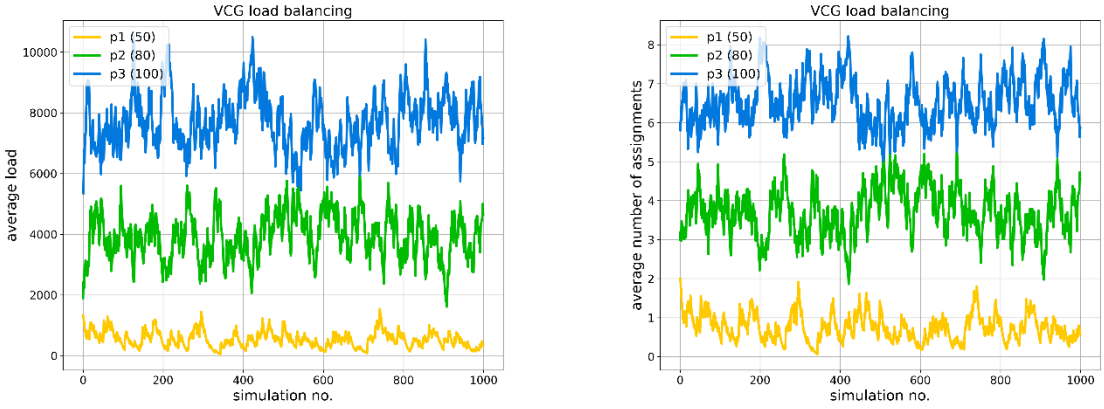
Mecanismul promovează licitarea sinceră prin compatibilitatea stimulentelor (licitarea sinceră este o strategie dominantă), raționalitatea individuală (fiecare agent sincer primește o plată), eficiența (licitarea sinceră maximizează bunăstarea socială).

Modelul de licitație VCG se bazează pe o implementare Python care primește datele de intrare dintr-un fișier text, descriind articolele și ofertele disponibile pentru fiecare participant și returnează alocarea articolelor în conformitate cu strategia de licitare veridică. Matricea distanțelor este tradusă în formatul de intrare după cum urmează: sumele licitate pentru fiecare sarcină sunt definite în funcție de distanță (costul raportat) și reputația normalizată a participantului. Beneficiul total și numărul de articole pentru fiecare participant sunt extrase pentru a compara rezultatele.

Folosind o metodă similară cu cea prezentată în cazul VRP, modelul VCG a fost integrat în simulator și rezultatele au fost evaluate pentru configurații aleatorii în ceea ce privește pozițiile de start ale participanților. Același număr de iterații (1000) a fost folosit pentru a compara rezultatele în ceea ce privește echilibrarea optimă a sarcinii într-o distribuție variabilă a participanților pe hartă.

Pentru inspecția vizuală, rezultatele sunt post-procesate folosind un filtru de medie alunecătoare (sliding window), așa cum se arată în Fig. 50 pentru fiecare tip de participant. Am luat în considerare aceleași tipuri de participanți cu reputație normalizată: 1 (raportare veridică și exactă), 0.8 (raportare veridică și în cea mai mare parte exactă) și 0.5 (raportare falsă/inexactă).

Rezultatele arată că atât un număr mai mare de sarcini, cât și o recompensă totală mai mare sunt atribuite pentru licitarea sinceră pe întreaga gamă de trasee simulate, care este direct influențată de reputația relativă a participanților.



(a) (b)

Fig. 50 Alocarea activităților prin VCG (a) Recompensă medie; (b) Sarcini alocate în medie.

5.4 Blockchain (B)

Componenta blockchain este construită pe baza tehnologiei Ethereum și a unora dintre cele mai noi platforme de dezvoltare, folosind Truffle, Ganache, Metamask și Web3.js.

5.4.1 Truffle

Pentru implementarea soluției blockchain bazată pe tehnologia Ethereum s-a folosit Truffle (ConsenSys Software Inc., 2021) ca mediu de dezvoltare și testare. Acesta oferă numeroase facilități pentru realizarea unei aplicații blockchain printre care: compilarea smart contract-urilor, testarea acestora, linie de comandă, pipeline-uri configurabile, interogarea rețelei blockchain pentru vizualizarea detaliilor fiecărei adrese (ETH disponibil pentru fiecare nod, cantitate deținută - moneda WGT), etc.

5.4.2 Ganache

Ganache[14] este o componentă a suitei Truffle și oferă posibilitatea realizării on-click a unei rețele blockchain locală pe baza căreia se pot dezvolta smart contracts. Prin intermediul Ganache avem acces atât la o interfață UI prin care se pot vedea în timp real wallet-urile / nodurile conectate la rețeaua locală, cât și la un utilitar de tip linie de comandă (Fig. 51).

Ganache poate fi folosit pe tot parcursul dezvoltării, pentru dezvoltare, deploy și testarea aplicației implementate.

[14] https://www.trufflesuite.com/docs/ganache/overview

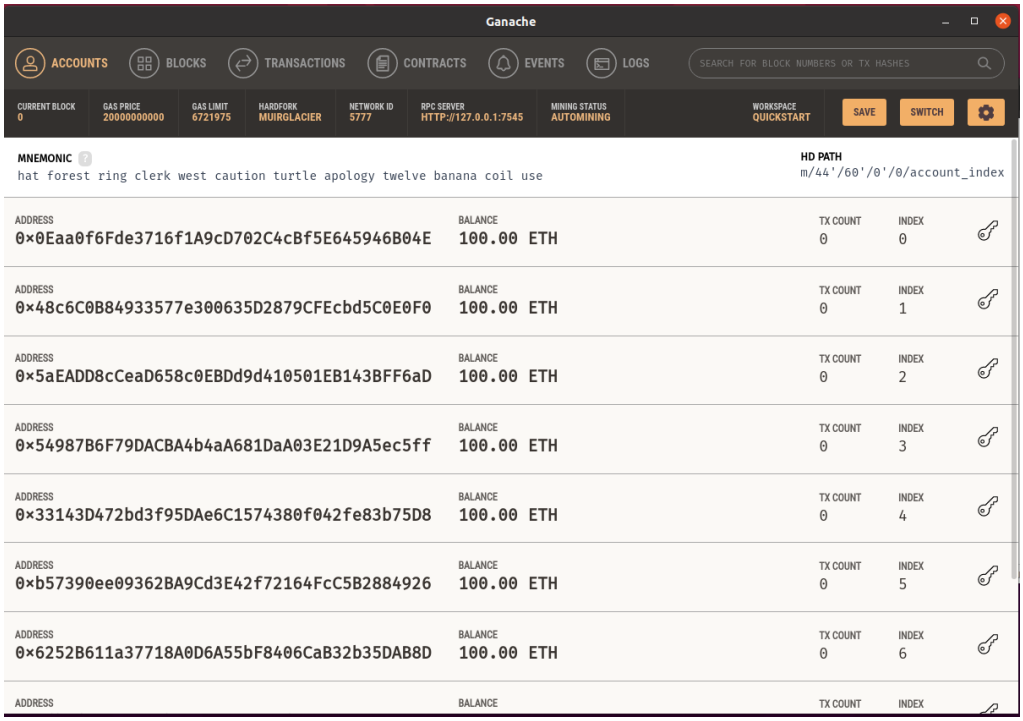


Fig. 51 Interfata Ganache

5.4.3 Smart Contracts

Pentru realizarea funcționalităților soluției blockchain este necesară implementarea contractelor inteligente. Astfel au fost dezvoltate următoarele: contract pentru realizarea monedei virtuale (WGT – Water Game Token), conform standardului ERC-20 (ethereum.org, 2021), monedă ce va fi oferită ca recompensa pentru clienți în urma acțiunilor acestora, dar care va putea fi și tranzacționată separat ca orice altă criptomonedă; contracte pentru facilitarea următoarelor acțiuni: cumpărare de WGT, transfer de WGT dintr-un portofel în altul, sau donare de WGT.

Implementarea contractelor inteligente a fost realizată folosind limbajul de programare Solidity[15], limbaj orientat pe obiect, specific pentru a interacționa cu tehnologia blockchain.

5.4.4 Metamask si Web3.js

Pentru facilitarea interacțiunii utilizatorului cu soluția blockchain s-a folosit Web3.js[16] împreună cu Metamask (MetaMask, 2021) (vezi Fig. 52). Metamask va fi folosit ca un gateway și oferă posibilitatea comunicării cu aplicații de tip blockchain, dar poate fi folosit și ca wallet. Metamask oferă posibilitatea de administra mai multe conturi / chei de conturi, de a tranzacționa criptomonede într-un mod securizat și a accesa aplicații descentralizate prin intermediul unei extensii ce poate fi instalată la nivelul unui browser web sau prin intermediul unei aplicații mobile.

[15] https://docs.soliditylang.org/en/v0.8.10/

[16] https://web3js.readthedocs.io/en/v1.5.2/

Web3.js reprezintă o colecție de biblioteci JavaScript prin intermediul cărora se facilitează interacțiunea cu un nod din rețeaua blockchain în scopul efectuării diferitelor operațiuni folosind o conexiune HTTP sau IPC (Inter-Process Communication).

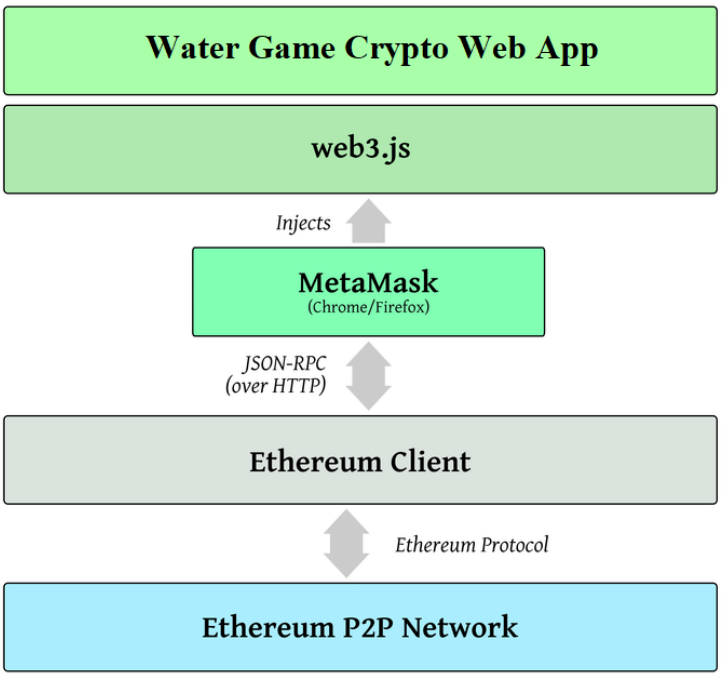


Fig. 52 Proiectarea componentei blockchain

6 Implementarea componentelor sistemului

Implementarea componentelor presupune metode și tehnologii eterogene, de la componenta hardware de colectare a datelor de consum, la sistemul colaborativ pentru interacțiunea cu utilizatorul, la sistemul de suport decizional și până la infrastructura blockchain pentru validarea tranzacțiilor de stimulare a participării. Arhitectura decuplată la nivel de sistem va permite integrarea acestor componente în următoarea etapă.

6.1 Monitorizare consum (WMS)

Implementarea componentei de monitorizare a consumului presupune conectarea debitmetrelor YF-S201 la modulele WiFi de tip NodeMCU (Karray et al., 2018). Majoritatea debitmetrelor disponibile comercial sunt debitmetre construite și care funcționează pe baza efectului Hall. Acest lucru determină ca în construcția senzorilor de măsurare din cadrul debitmetrelor să existe un senzor magnetic de efect Hall integrat care, folosind o rotiță, produce un impuls electric la fiecare rotație a acesteia. Astfel valoarea primară pe care un debitmetru o furnizează este frecvența impulsurilor sau numărul de impulsuri pe unitatea de timp. Pe baza acestei valori a frecvenței impulsurilor se calculează valoarea debitului de apă în litri / minut sau litri pe oră. Formula de calcul utilizată pentru modelele de debitmetre de acest tip presupune o împărțire a valorii frecvenței impulsurilor la un factor numeric supraunitar.

În fiecare locuință, au fost instalate debitmetre și un număr minim de module WiFi în funcție de configurația locală. În general, un singur modul a fost instalat în fiecare locație (e.g. baie, bucătărie), la care au fost conectate debitmetrele prin cabluri (de ex.: chiuvetă apă caldă / rece, duș combinat, toaletă, mașină de spălat).

6.1.1 Componentele de monitorizare debit de apă și nivel edge (UTCN.C)

Pentru sistemul de monitorizare a debitului de apă s-au realizat 3 modele de montaj.

În primul caz (UTCN.C1) s-au folosit în locații debitmetre de tipul YF-S201 conectate la noduri NodeMCU cu chip wireless de tip ESP8266[17]. Alimentarea acestora la punctele de măsurare s-a făcut cu surse de tip AC. Aceste surse au fost înlocuite în situațiile în care spațiul disponibil pentru montaj și sursele de alimentare cu curent alternativ nu sunt posibile, cu surse de alimentare bazate pe acumulatori externi conform montajului din Fig. 53. Acest tip de alimentare este posibilă doar pe perioade limitate de timp, cu necesitatea înlocuirii sursei de alimentare o dată la 96 ore. Mai departe acestea au fost configurate să se conecteze la punctul local de acces din fiecare locație unde sunt instalate, pentru a putea comunica cu serviciul web din cloud și pentru a transmite acestuia datele de consum.

La nivelul nodului NodeMCU se utilizează un sketch care programează partea de configurare a nodului inclusiv din punctul de vedere al conectivității, execută partea de preprocesare a datelor, calculează consumul de apă, iar apoi face postarea datelor formatate cu câmpurile de identificare spre serviciul din cloud.

[17] https://usermanual.wiki/Document/YFS201manual.662398285/view

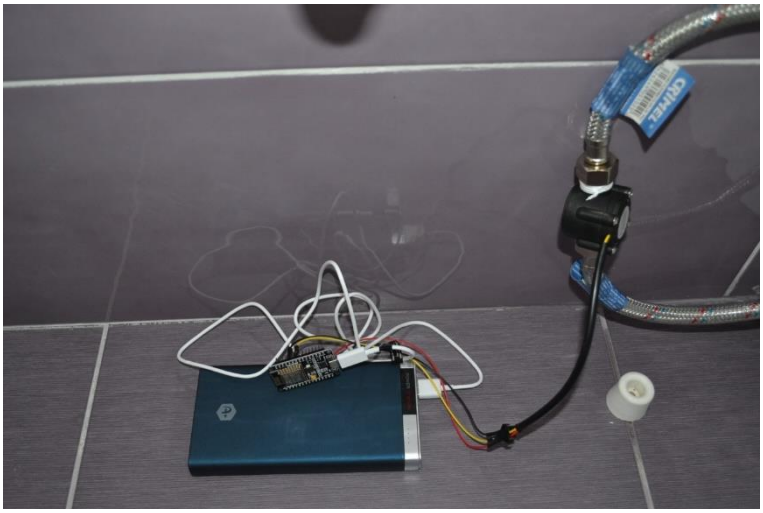


Fig. 53 Montaj debitmetru conectat la NodeMCU alimentat cu sursă externă

Modelul de sketch folosit local pe nodul NodeMCU pentru testarea modului de calcul a debitului de apă este următorul:

```
void setup()
{
   pinMode(flowmeter, INPUT);
   Serial.begin(9600);
   attachInterrupt(0, flow, RISING);
   sei();
   currentTime = millis();
   cloopTime = currentTime;
}
void loop ()
{
   currentTime = millis();
   if(currentTime >= (cloopTime + 1000))
   {
      cloopTime = currentTime;
      l_hour = (flow_frequency * 60 / 7.5);
      flow_frequency = 0;
      Serial.print(l_hour, DEC);
      Serial.println(" L/hour");
   }
}
```

Conform specificațiilor producătorului debitmetrului YF-S201, formula de calcul pentru debitul de apă măsurat în litri / oră este: Debit (l / h) = Frecvența_impulsurilor * 60 / 7.5.

În al doilea caz (UTCN.C2) sunt montate multiple debitmetre la mai multe noduri NodeMCU aflate în puncte diferite în cadrul aceleiași locații. Tot în cadrul locației este instalat un SBC de tip Raspberry PI Model 3+[18]. În Fig. 54 se poate observa un montaj cu debitmetru și NodeMCU alimentat la sursa de tip AC / DC.

Fig. 54 Montaj debitmetru conectat la NodeMCU alimentat cu sursa AC / DC

În acest caz, la nivelul nodului NodeMCU s-a testat funcționarea comunicării cu SBC-ul Raspberry PI prin intermediul sketch-ului următor:

```
unsigned long timerDelay = 5000;

void setup() {

Serial.begin(9600);

WiFi.begin(ssid, password);

Serial.println("Connecting");

      while(WiFi.status() != WL_CONNECTED) {

          delay(500);

          Serial.print(".");

        }
```

[18] https://static.raspberrypi.org/files/product-briefs/Raspberry-Pi-Model-Bplus-Product-Brief.pdf

```
        Serial.println("");
        Serial.print("Connected to WiFi network with IP Address: ");
        Serial.println(WiFi.localIP());
      }
void loop() {
      if ((millis() - lastTime) > timerDelay) {
        if(WiFi.status()== WL_CONNECTED){
          HTTPClient http;
          http.begin(serverName);
          http.addHeader("Content-Type","application/x-www-form-urlencoded");
          http.addHeader("Content-Type", "application/json");
          int httpResponseCode = http.POST("\"12\"");
          Serial.print("HTTP Response code: ");
          Serial.println(httpResponseCode);
          http.end();
          }
          else {
            Serial.println("WiFi Disconnected");
          }
          lastTime = millis();
        }
      }
```

La nivelul nodurilor NodeMCU se calculează, bazat pe valoarea frecvenței impulsurilor, doar consumul în litri / oră, care este transmis formatat cu datele de identificare ale punctului de măsurare către SBC. La nivelul SBC-ului se face pre-procesarea datelor de consum, care ulterior sunt transmise în cloud pentru vizualizare și analiză mai avansată.

În cazul al treilea (UTCN.C3), similar cu ceea ce se întâmplă în cazul 2, se conectează debitmetrele la nodurile NodeMCU corespunzătoare în funcție de dispunerea punctelor de măsurare în cadrul locației după care acestea execută citirile datelor de consum, calculează consumul în litri / minut, pre-procesează datele și le transmit mai departe unui dispozitiv embedded cum este cel de la Nvidia Jetson Nano. La nivelul acestui dispozitiv se face partea de procesare a datelor.

6.1.2 Serviciul de colectare a datelor (UTCN.C)

Soluția aleasă pentru serviciul de colectare a datelor este o implementare open source a standardului OGC SensorThings, Frost server[19]. Implementarea este una bazată pe containere

[19] https://fraunhoferiosb.github.io/FROST-Server/

Docker. Aplicația conține două containere, unul pentru serviciul de colectare a datelor și altul pentru baza de date PostgreSQL cu extensie Postgis. Aplicația poate primi atât cereri HTTP cât și cereri MQTT. În prezent, componentele sistemului dezvoltat în proiect utilizează doar cereri HTTP.

Pentru instalare în cloud s-a folosit următorul fișier Docker compose:

```
version: '2'


services:
  web:
    image: fraunhoferiosb/frost-server:latest
    environment:
      - serviceRootUrl=http://10.20.1.47/:8080/FROST-Server
      - http_cors_enable=true
      - http_cors_allowed.origins=*
      - persistence_db_driver=org.postgresql.Driver
      - persistence_db_url=jdbc:postgresql://database:5432/sensorthings
      - persistence_db_username=sensorthings
      - persistence_db_password=xxx
      - persistence_autoUpdateDatabase=true
    ports:
      - 8080:8080
      - 1883:1883
    depends_on:
      - database


  database:
    image: postgis/postgis:11-2.5-alpine
    environment:
      - POSTGRES_DB=sensorthings
      - POSTGRES_USER=sensorthings
      - POSTGRES_PASSWORD=xxx
    volumes:
      - postgis_volume:/var/lib/postgresql/data
volumes:
    postgis_volume:
```

În Fig. 55 se poate observa răspunsul serverului Frost la o interogare despre Observații:

Fig. 55 Răspunsul serverului Frost la o interogare despre Observații

API-ul Frost poate să răspundă la interogări legate de toate entitățile specificate în modelul de date (Senzori, Locații, Datastreams etc.).

6.1.3 Configurarea modulelor de achiziție (UPB.C)

Modulele NodeMCU (NodeMcu Team, 2018) sunt inițial programate în Arduino IDE și expun anumite configurări necesare pentru instalarea în locația consumatorului (Bento, 2018).

Configurarea plăcilor se realizează prin intermediul WiFi, acestea funcționând atât în modul de stație, cât și în modul AP (access point). Pagina de configurare a modulelor este găzduită local și poate fi accesată prin conectarea la punctul de acces propriu la adresa 192.168.4.1 (Fig. 56). Conexiunea este securizată la nivel de protocol WiFi, prin WPA-PSK (Khasawneh et al., 2014).

Se inițializează parametrii precum: configurația pinilor la care sunt conectate debitmetrele și parametrii de conectare (numele și parola rețelei WiFi locale). Pagina de configurare oferă și informații suplimentare legate de măsurătorile realizate ușurând astfel instalarea și depanarea la fața locului.

Configurările sunt salvate în memoria EEPROM și pot fi modificate doar prin conectarea la modul, caz în care este solicitată parola de WiFi, sau prin comenzi de la distanță, prin intermediul mesajelor transmise prin MQTT, caz în care este solicitată o parolă pentru conectare.

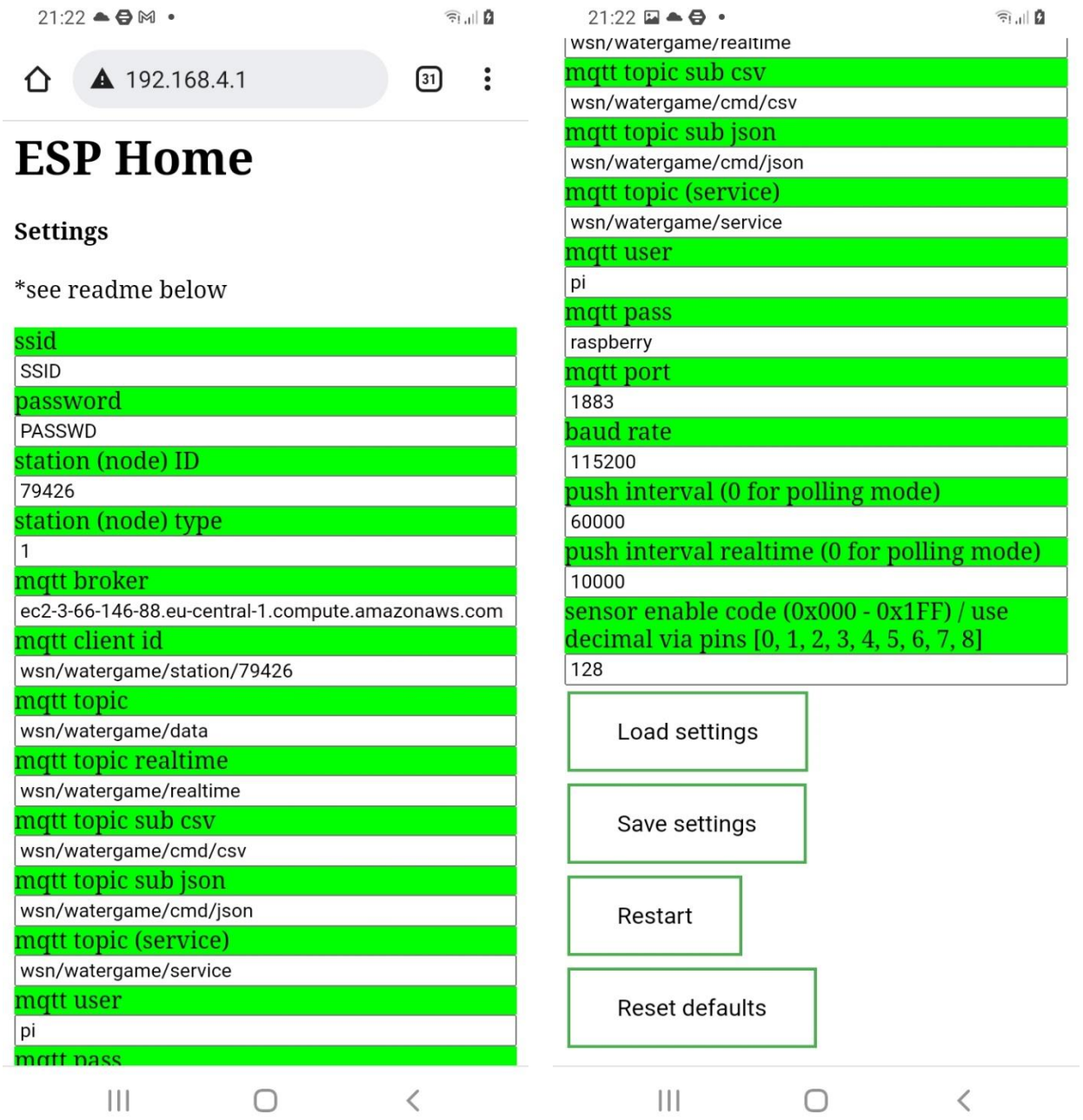


Fig. 56 Interfața integrată de configurare a modulului de achiziție

6.1.4 Interfață de comunicare (UPB.WSN)

Interfața de comunicare este proiectată pentru a facilita transmiterea datelor în rețeaua de senzori wireless (WSN). Prin intermediul conexiunii WiFi locale, modulele se conectează la brokerul MQTT din Cloud și trimit pachete de date la diferite intervale:

- date de debit - frecvență ridicată (5 s), pe topicul wsn/watergame/realtime;
- date de volum - frecvență scăzută (10 min), pe topicul wsn/watergame/data;
- date de control - configurare de la distanță, pe topicul wsn/watergame/cmd.

Datele sunt colectate de către server, acesta fiind abonat la brokerul de MQTT, astfel:

- Datele cu frecvență redusă sunt stocate în baza de date MySQL pentru păstrarea istoricului de consum pe termen lung.
- Datele cu frecvență ridicată sunt utilizate pentru afișarea în timp real și sunt stocate temporar în memoria REDIS.

Prin aplicația client MQTTBox[20] se pot vizualiza datele care se transmit în timp real. Atunci când modulele trimit datele către brokerul de MQTT, se realizează operația de "publish" ce poate fi monitorizată prin abonarea la acel topic (Fig. 57). Atunci când comenzile sunt recepționate de module prin brokerul de MQTT, se realizează operația de "subscribe", iar comenzile pot fi trimise prin operația de “publish” din aplicația client (Mishra & Kertesz, 2020).

Pot exista astfel mai mulți abonați, precum server-ul NodeJS aflat în cloud pe Amazon EC2 (Amazon Web Services, 2021), care preia datele și le stochează în baza de date MySQL sau aplicația mobilă ce afișează datele în timp real.

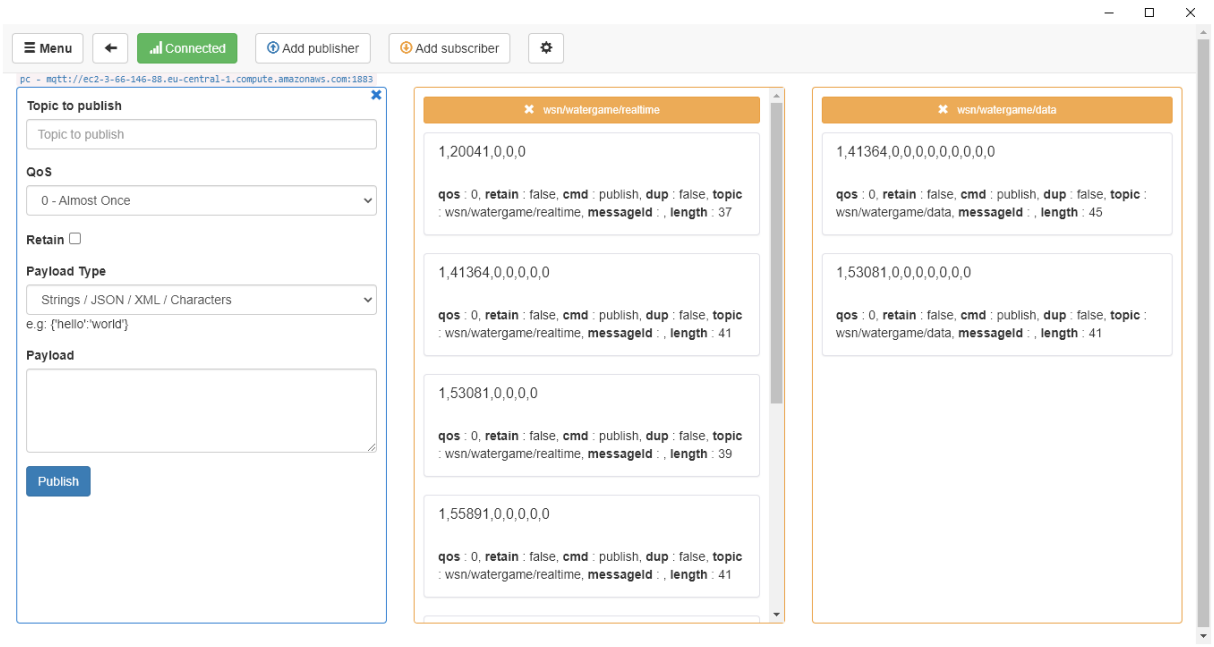


Fig. 57 Vizualizarea datelor cu MQTTBox

Pentru identificarea senzorilor, datele trimise de un modul sunt împachetate în format CSV (comma-separated values), având următoarea structură: "node_type, node_id, recv_command, channel_0_value, channel_1_value, ...", unde:

- **node_type** - tipul elementului care a trimis datele (în acest caz avem doar senzori de debit, având codificarea 1);
- **node_id** - id-ul modulului de achiziție;
- **recv_command** - tipul comenzii (100 pentru identificarea pachetului de date, 110 pentru identificarea comenzilor);

[20] https://chrome.google.com/webstore/detail/mqttbox/kaajoficamnjijhkeomgfljpicifbkaf

- **channel_0_value, channel_1_value, …** - valorile în format CSV corespunzătoare canalelor de măsurare configurate pentru modulul de achiziție.

6.1.5 Procesul de instalare (UPB)

Instalarea constă în montarea debitmetrelor și a modulelor de comunicație. Debitmetrele se instalează prin racord flexibil de 1/2” cu garnitură de etanșare. Pentru situațiile în care se folosește racord de 3/8” se adaugă un racord sau adaptor 1/2” la 3/8”. În cazul bateriei de la duș, montarea senzorului se realizează prin racord fix de 1/2” iar furtunul de la duș se conectează direct la celălalt capăt al debitmetrului (Fig. 58).

Modulele de achiziție sunt amplasate în locații mai greu accesibile, precum masca de la chiuvetă. Debitmetrele se conectează prin cabluri la un modul de achiziție, fapt pentru care este necesară o evaluare a locației în vederea dimensionării și amplasării corespunzătoare.

Configurarea modulelor se realizează după ce au fost montate debitmetrele, prin conectarea la pagina de configurare, apoi se verifică datele colectate.

În următoarele imagini sunt prezentate câteva module instalate în locuințe. În Fig. 59 este prezentat un debitmetru conectat la bateria de duș. Cablurile sunt instalate într-un traseu de cabluri pentru o conexiune robustă. Distanța dintre module și debitmetre nu a reprezentat un impediment pentru captarea impulsurilor generate de debitmetre, fiind instalate debitmetre la distanțe de la 20 cm până la 10 m.

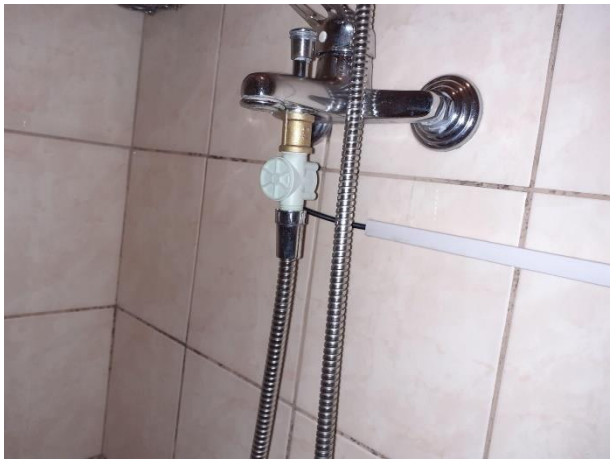


Fig. 58 Senzor de debit instalat la bateria de duș

În Fig. 59 este prezentat un modul de achiziție montat sub masca de la chiuvetă și conexiunile cu debitmetrele atașate la alimentarea cu apă caldă și rece.

Fig. 59 Modul de achiziție instalat sub chiuvetă

În Fig. 60 este prezentat un debitmetru instalat la alimentarea cu apă a toaletei, conectat prin cabluri la modulul de achiziție aflat pe peretele opus, sub masca de la chiuvetă. Cablurile sunt instalate prin traseul de cabluri atașat de perete.

Fig. 60 Senzor de debit instalat la alimentarea cu apă a toaletei

În primă fază, am instalat 32 de senzori de debit conectați la 11 module de achiziție conform tabelului de mai jos:

Tabel 2 Module de achiziție instalate pentru colectarea datelor de consum din locuințe

modul				D1	D2	D3	D5	D6	D7	localizare	
no.	id	locație	spațiu	chiuvetă rece	chiuvetă caldă	toaletă	duș	mașină spălat	mașină spălat vase	lat	lng
1	97074	DA	bucătărie	x	x					44.48634	26.04781
2	20041	CC	bucătărie	x	x					44.411789	26.1273017

3	62706	CC	baie_1	x	x	x				44.411789	26.1273017
4	41364	CC	baie_2	x	x	x	x			44.411789	26.1273017
5	33383	CC	baie_3	x	x	x	x			44.411789	26.1273017
6	53081	CC	baie_4	x	x	x				44.411789	26.1273017
7	55891	CC	baie_5	x	x	x	x			44.411789	26.1273017
8	16102	AT	bucătărie	x	x					44.414563	26.0335787
9	14945	AT	baie			x				44.414563	26.0335787
10	89776	AB	bucătărie	x	x					44.377512	26.1381178
11	65506	AB	baie	x	x	x		x	x	44.377512	26.1381178

Fiecare modul este reprezentat printr-un ID unic generat la instalare și este configurat pentru colectarea datelor de la senzorii de debit atașați. Astfel, pentru fiecare intrare de date (canal de măsurare), au fost atașați senzorii conform tabelului: primele 2 canale (D1, D2) sunt folosite pentru chiuvetă (apă rece, apă caldă), al treilea (D3) pentru toaletă, al cincilea (D5) pentru duș, iar ultimele 2 canale (D6, D7) pentru mașina de spălat, respectiv mașina de spălat vase.

Modulele de achiziție au fost programate anterior instalării și permit configurarea intrărilor de date din interfața de configurare. În urma testelor, pinul D4 a cauzat probleme de stabilitate, fiind identificată ulterior corespondența cu funcția de inițializare a modulului ESP8266 în modul de programare. Astfel, pinul D4 nu a mai fost conectat, iar senzorul pentru duș a fost conectat la pinul D5.

6.1.6 Aplicație mobilă (UPB.WMS)

Aplicația mobilă WaterMonitoringSystem (WMS) disponibilă pentru Android permite monitorizarea consumului de apă din perspectiva consumatorilor sau a furnizorilor, oferind o interfață interactivă pentru sistemul de monitorizare a consumului (Fig. 61). Datele de consum sunt preluate de la server-ul aplicației, iar autentificarea se realizează prin Firebase pentru o securitate sporită și decuplarea serviciilor de date (datele colectate de la senzori sunt gestionate separat de cele referitoare la conturile utilizatorilor).

Aplicația mobilă implementează cazurile de utilizare identificate în etapa de proiectare a arhitecturii sistemului, precum: vizualizarea senzorilor pe hartă, vizualizarea graficelor, cont furnizori / consumatori, raportare probleme.

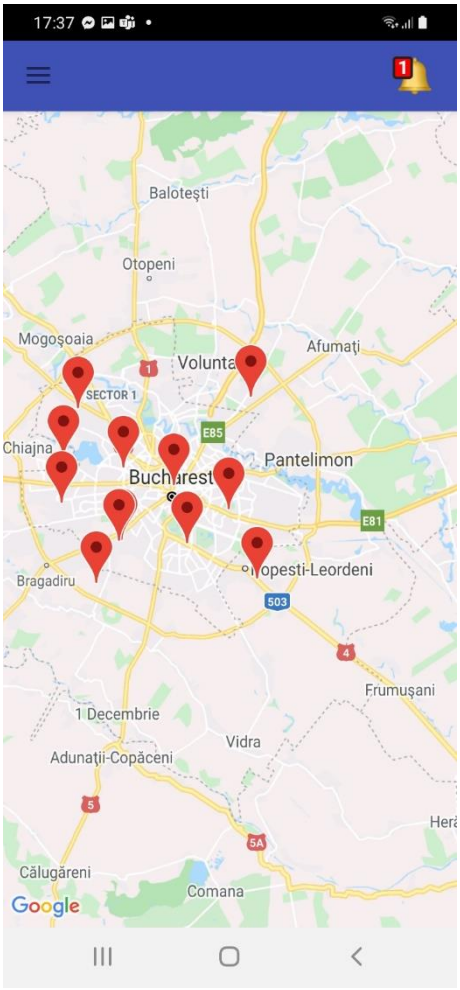


Fig. 61 Aplicația WaterMonitoringSystem (mod furnizor). Vizualizarea pe hartă a modulelor instalate.

Implementarea aplicației este realizată în Android Studio, folosind limbajul Java. Pentru conectarea la serviciul Firebase, se configurează un proiect în Firebase Console și o aplicație mobilă (Khawas & Shah, 2018). Platforma generează un fișier de configurare care este adăugat în proiectul Android, preluat de SDK-ul Firebase pentru conectare la baza de date.

Aplicația este formată din următoarele pachete Java:

- **activities** – pachetul în care se găsesc clasele responsabile pentru ferestrele aplicației, structurate în funcție de rol;
- **adapters** – clase responsabile cu funcționalități mai complexe, precum actualizarea listelor cu date în timp real;
- **api** – clase responsabile de conectarea la server-ul NodeJS prin cereri HTTP;
- **authentication**: clase responsabile cu procesul de autentificare și înregistrare a utilizatorilor în aplicație;
- **models** – clase ce modelează datele și mapările necesare pentru interacțiunea cu server-ul aplicației și cu serviciul Firebase;

- **mqtt** – clase ce implementează comunicația prin MQTT folosind biblioteca Paho[21] și un set de metode pentru conectarea la fluxul de date în timp real;
- **utils** – funcții de suport pentru operații mai complexe.

6.1.7 Serviciul de date (UPB.REST)

Serviciul de date este implementat pe back end-ul principal al sistemului de monitorizare a consumului. Aplicația mobilă folosește serviciile disponibile sub formă de REST API în cadrul server-ului principal (Masse, 2011). Acesta expune un set de metode pentru accesarea datelor colectate și a informațiilor despre modulele de achiziție înregistrate conform tabelului de mai jos:

Tabel 3 REST API

endpoint	**descriere**	**parametri**
GET /sensors/manager/get-registered-sensors	vizualizarea modulelor înregistrate în baza de date	N/A
GET sensors/data/get-current-data-snapshot	vizualizarea ultimelor date transmise de un anumit modul	sensorId
GET sensors/data/get-stored-data	vizualizarea datelor înregistrate în baza de date, în format JSON	sensorId chan limit mode
GET /sensors/data/get-stored-data-processed	vizualizarea datelor grupate după ID-ul modulului de achiziție, în format JSON	sensorId chan limit mode
GET /sensors/data/get-stored-data-csv	vizualizarea datelor grupate după ID-ul modulului de achiziție, în format CSV	sensorId chan limit mode
POST /sensors/manager/set-coords	asocierea coordonatelor GPS pentru un modul de achiziție	sensorId lat lng

În următorul tabel este prezentat modul în care endpoint-urile sunt corelate cu funcționalitățile aplicației:

[21] https://www.eclipse.org/paho/index.php?page=downloads.php

Tabel 4 Utilizarea endpoint-urilor în aplicație

endpoint	descriere	funcționalitate
GET /sensors/manager/get-registered-sensors	vizualizarea modulelor înregistrate în baza de date	afișarea pe hartă a modulelor instalate
GET sensors/data/get-current-data-snapshot	vizualizarea ultimelor date transmise de un anumit modul	inițializarea datelor afișate pentru fiecare senzor
GET sensors/data/get-stored-data	vizualizarea datelor înregistrate în baza de date, în format JSON	afișarea datelor colectate de un modul de achiziție în format grafic (un singur senzor)
GET /sensors/data/get-stored-data-processed	vizualizarea datelor grupate după ID-ul modulului de achiziție, în format JSON	afișarea datelor colectate de un modul de achiziție în format grafic (mai mulți senzori)
GET /sensors/data/get-stored-data-csv	vizualizarea datelor grupate după ID-ul modulului de achiziție, în format CSV	export date în fișier CSV
POST /sensors/manager/set-coords	asocierea coordonatelor GPS pentru un modul de achiziție	adăugarea sau actualizarea coordonatelor GPS pentru un modul de achiziție

6.2 Suport decizional (DSS)

6.2.1 Serviciul de procesare/analiză a datelor (UTCN.A)

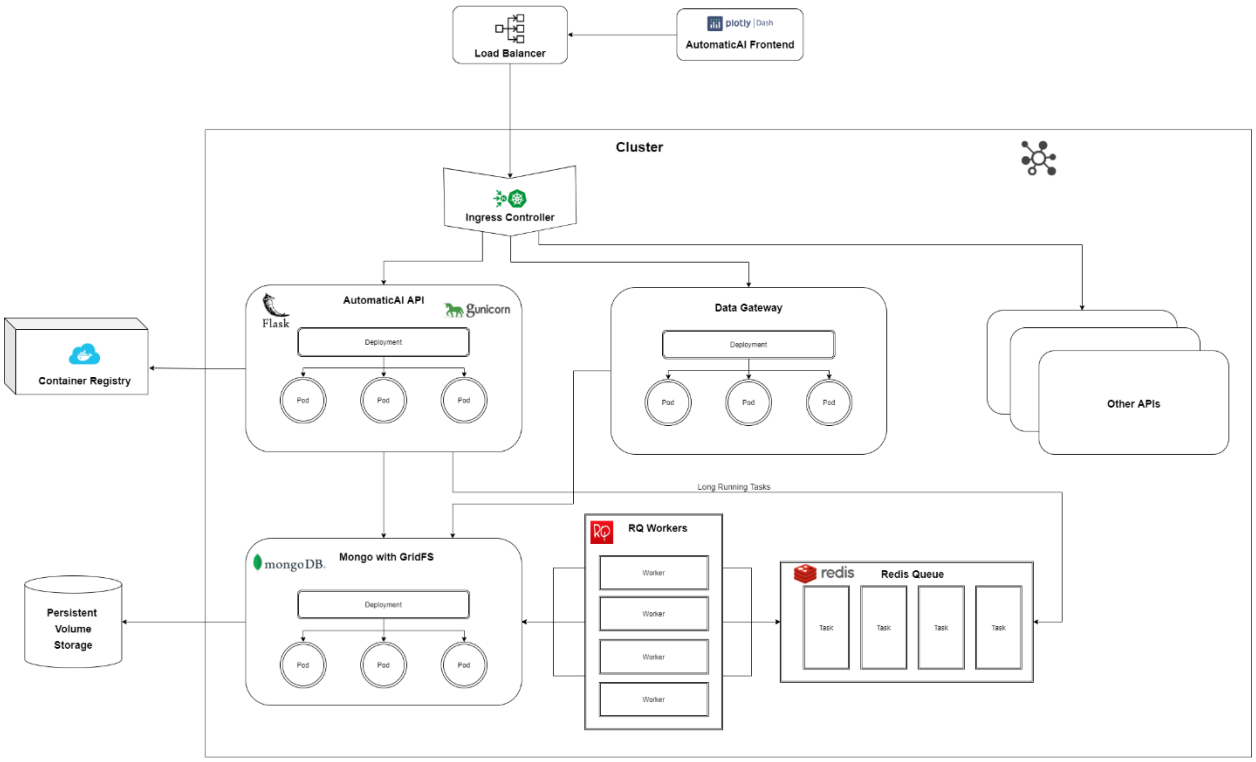


Fig. 62 Arhitectura detaliată a platformei extinse AutomaticAI

După cum se poate vedea în Fig. 62, există un site web implementat în Dash Plotly care comunică cu aplicația bazată pe microservicii printr-un load balancer. Folosind acest load balancer, se decuplează clientul de clustere și se expune un IP public prin care serviciile din cluster pot fi apelate de clienți. Deoarece request-urile diferite pot ajunge la diferite servicii pe diferite noduri sau mașini virtuale, a fost folosit un controler de intrare (ingress controller) NGINX, care funcționează ca un proxy invers ascunzând diferitele mașini virtuale de la client și, de asemenea, rutează cererile pe baza sub-domeniului furnizat în adresa URL a request-ului.

În Fig. 62, unul dintre cele mai importante microservicii este AutomaticAI API. Acest serviciu conține logica pentru configurarea unui pipeline pentru rezolvarea problemelor de clasificare, regresie sau de detectare a anomaliilor. Pipeline-ul poate fi configurat pe baza unui json de configurare. Acest serviciu conține, de asemenea, algoritmul PSO-SA, care poate selecta automat un tip de algoritm AI și poate regla hiper-parametrii acestuia doar pe baza datelor de intrare, fără intervenție umană.

Selectarea, reglarea și antrenarea unui algoritm AI este un proces consumator de timp și de lungă durată. Pentru a preveni blocarea API-ului, se rulează toate procesele de lungă durată în paralel și asincron. Pentru a obține o paralelizare adevărată și o procesare asincronă, a fost aplicat modelul producător-consumator, în care se folosește o coadă Redis pentru a programa sarcinile care trebuie procesate offline de către procesele de lucru numite workers.

Pentru stocarea fișierelor de date de intrare și a modelelor antrenate se folosește MongoDB cu GridFS. GridFS este utilizat pentru stocarea și preluarea fișierelor mari. Aceste fișiere sunt împărțite în bucăți și fiecare bucată este stocată ca document separat. De asemenea, se stochează metadate despre bucăți, astfel fișierul original putând fi reconstruit oricând.

Serviciul numit AutomaticAI API are structura prezentată în Fig. 63. El are mai multe module, fiecare având responsabilități diferite. Modulul principal este numit App, care face legătura între diferitele componente și module ale aplicației și definește rutele prin care request-urile pot ajunge la componenta care poate să le trateze.

Aplicația conține mai multe controllere:

1) UsersAPI – conține endpointuri de CRUD pentru utilizatori (adică metode de GET, POST, PUT și DELETE);
2) DatasetsAPI – conține endpointuri de CRUD pentru seturi de date;
3) AIModelAPI – endpoint pentru a citi modelul antrenat și salvat în baza de date;
4) TrainAPI – endpoint pentru declanșarea antrenării modelului selectat;
5) JobsAPI – pentru procesele de lungă durată se generează un job care se rulează în spate. Acest modul este folosit pentru a vizualiza job-urile și statusurile lor (e util pentru depanare);
6) ModelSelectionAPI – conține un endpoint care declanșează rularea algoritmului de selecție și optimizare automată a modelelor AI, numit PSO-SA;
7) PipelineAPI- modulul cel mai complex, folosit pentru a crea un pipeline configurabil, așa cum este descris în secțiunea următoare;
8) LstmAPI – conține un endpoint pentru crearea și configurarea unui model LSTM;
9) AutoArimaAPI - conține un endpoint pentru crearea și configurarea algoritmului Auto-Arima;
10) EvaluateAPI – conține diferite funcții pentru evaluarea algoritmilor de clasificare (acuratețe, precizie, scor f1, matrice de confuzie etc.) și de regresie (MAE, RMSE etc.);
11) TimeSeriesEvaluationAPI – conține diferite funcții pentru evaluarea algoritmilor folosiți pentru analiza, procesarea seriilor de timp și de predicție a valorilor următoare.

Fig. 63 Componentele serviciului AutomaticAI API

Pipeline dinamic

Pipeline-ul dinamic se configurează folosind un json de configurare. În mod obligatoriu, pipeline-ul trebuie să aibă un nume, trebuie să se specifice numele modelului care va fi salvat în urma rulării acestui pipeline, un nume de utilizator (astfel fiecare utilizator va avea acces doar la pipeline-urile create de el), numele fișierului de intrare, folosit ca date de antrenare, cantitatea datelor folosite din acest fișier exprimat în procente, și elementele pipeline-ului. Proprietatea ”pipelineElements” e o listă care conține niște obiecte având ca proprietăți tipul algoritmului și

parametrii. Ca să se poată crea o listă dinamică de parametri a fost utilizat un dicționar, unde cheia este numele parametrilor. Un exemplu pentru configurarea unui pipeline care conține un pas de scalare folosind algoritmul MinMax și un pas de re-echilibrarea datelor folosind algoritmul SMOTE (Chawla et al., 2002) este prezentat în Fig. 64.

```
{
    "name": "test-pipeline",
    "modelName": "test-model",
    "step": "first-step",
    "username": "test1",
    "trainDatasetName": "test.csv",
    "sampleSizeInPercentage": 100,
    "pipelineElement":
    [
        {
            "algorithmType": "Scale",
            "parameters":
            {
                "scaleType": "MinMax"
            }
        },
        {
            "algorithmType": "Rebalance",
            "parameters":
            {
                "rebalanceType": "SMOTE"
            }
        }
    ]
}
```

Fig. 64 Exemplu pentru configurarea pipeline-ului dinamic

Interfața grafică de tip web

Această interfață ajută în utilizarea mai ușoară a sistemului de procesare și analiză de date (Fig. 65). Setul de date de analizat poate fi încărcat în platformă prin drag-and-drop sau prin selecția fișierului în fereastra browse.

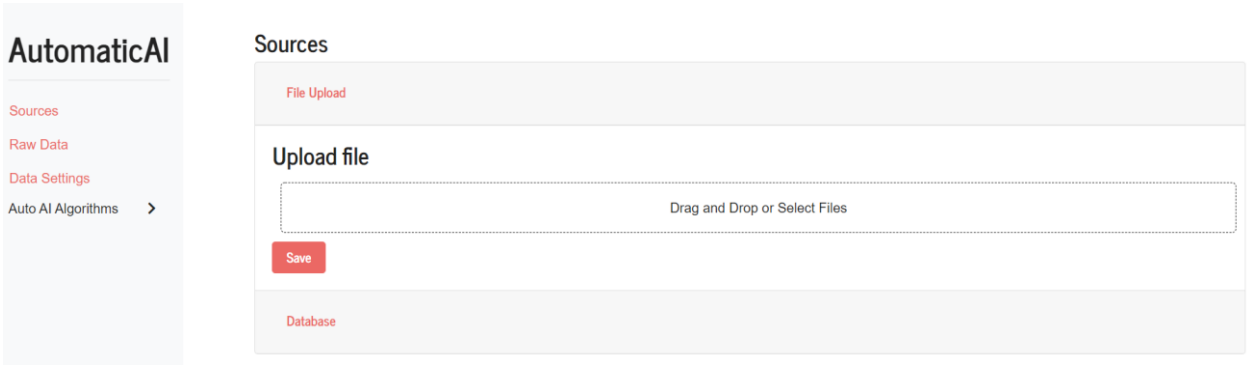


Fig. 65 Încărcarea fișierelor

În acest moment platforma suportă doar fișiere de tip .csv. După încărcarea fișierului, în secțiunea Raw Data se pot vizualiza toate fișierele încărcate de către un utilizator (Fig. 66).

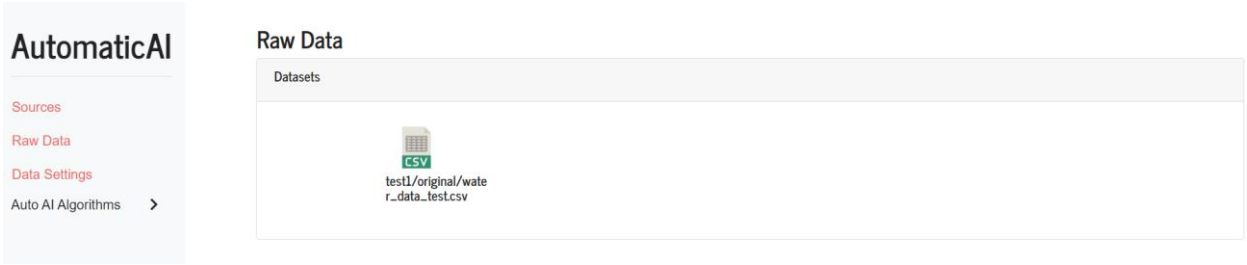


Fig. 66 Vizualizarea listei fișierelor încărcate

Pentru a specifica care este coloana țintă sau coloana care conține etichetele (în cazul problemelor de clasificare) există pagina de Data Settings, unde utilizatorul poate selecta coloana dorită, apăsând butonul numit Set as target (Fig. 67).

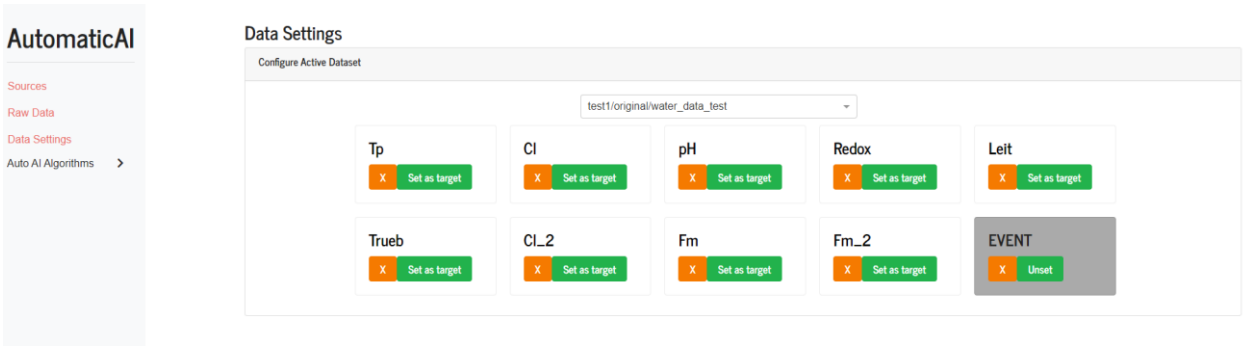


Fig. 67 Configurarea setului de date

După configurarea datelor de intrare, urmează crearea pipeline-ului dinamic. Primul pas este selecția datelor de intrare utilizând drop-down-ul din Fig. 68.

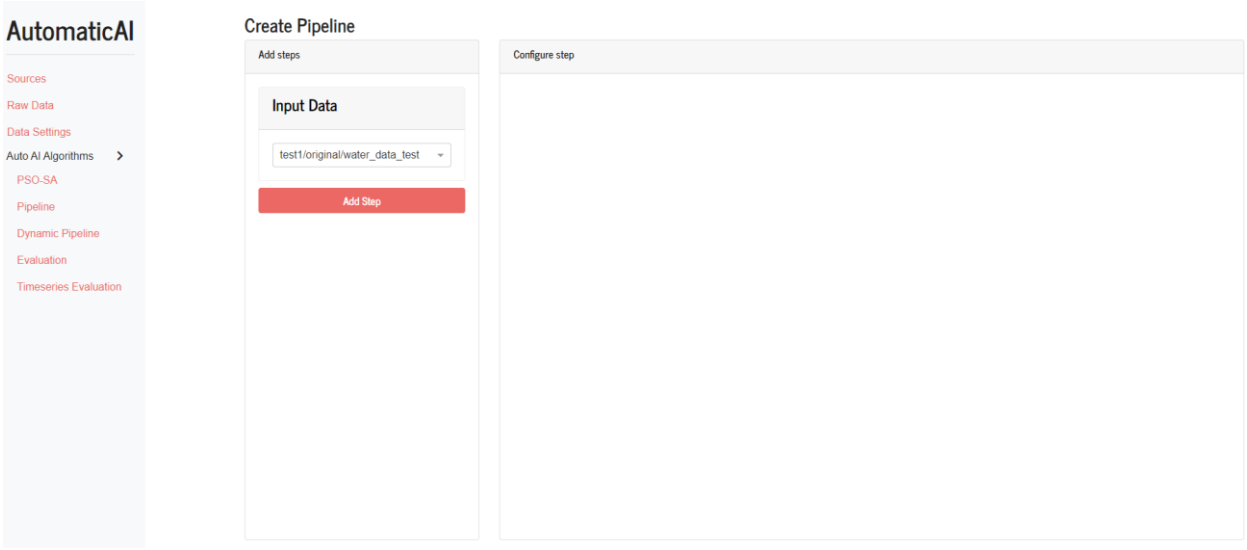


Fig. 68 Alegerea setului de date în configurarea pipeline-ului dinamic

Pentru a adăuga diferiți pași de pre-procesare, se poate apăsa butonul Add Step, apoi se alege categoria (cum ar fi transformarea datelor, scalare, normalizare etc.) și numele algoritmului (de exemplu MinMax, SMOTE, RUID etc.) (Fig. 69).

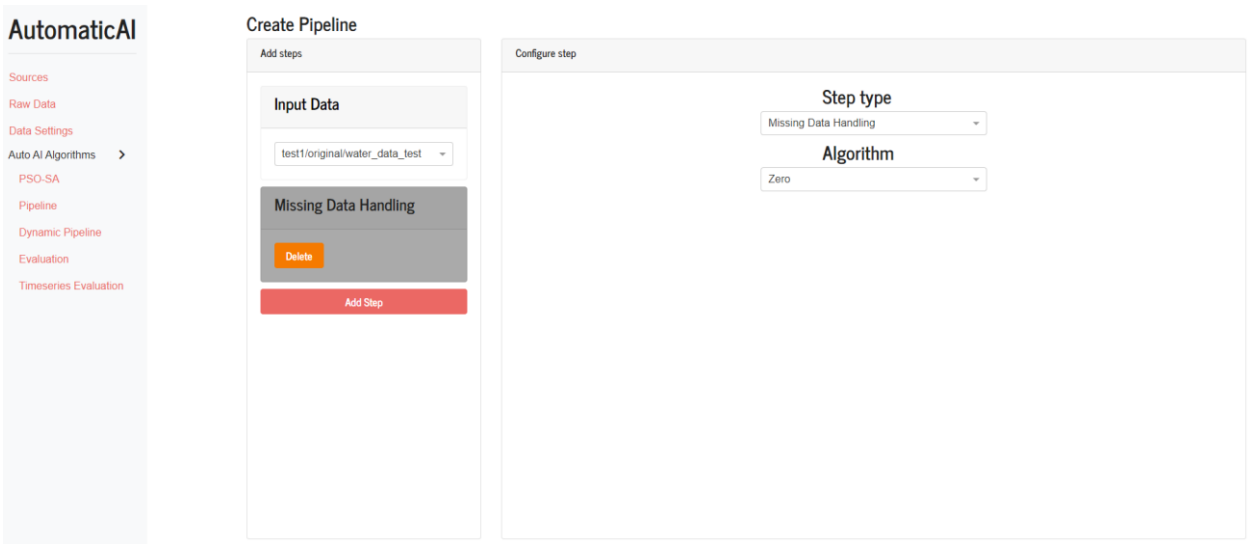


Fig. 69 Adăugarea unui pas de preprocesare în pipeline-ul dinamic

Ultimul pas este selecția algoritmului de analiză a datelor, cum ar fi algoritmi de clasificare, de analiză de serii de timp etc. Pentru fiecare algoritm se deschide un panou de control cu parametrii specifici algoritmului. Prin apăsarea butonului de Train începe procesul de antrenare, progresul fiind afișat folosind un progress bar (Fig. 70).

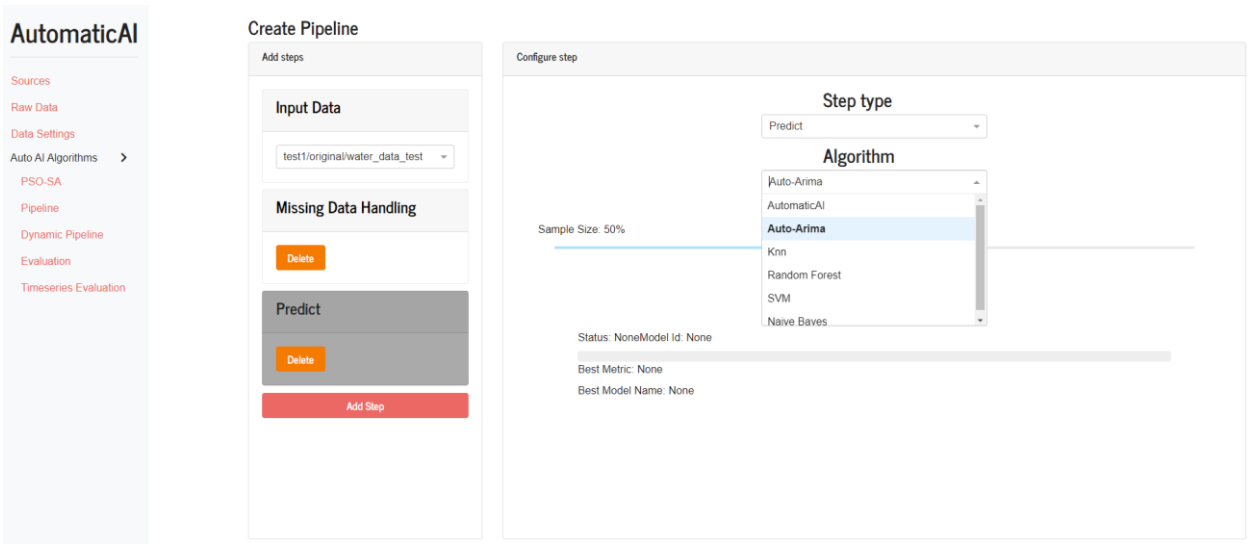


Fig. 70 Alegerea algoritmului de analiză de date

După antrenare, modelul poate fi aplicat pe date noi, pe date de test și automat se generează rapoarte și metrici de evaluare (cum ar fi precizie, acuratețe, matricea de confuzie etc.) (Fig. 71).

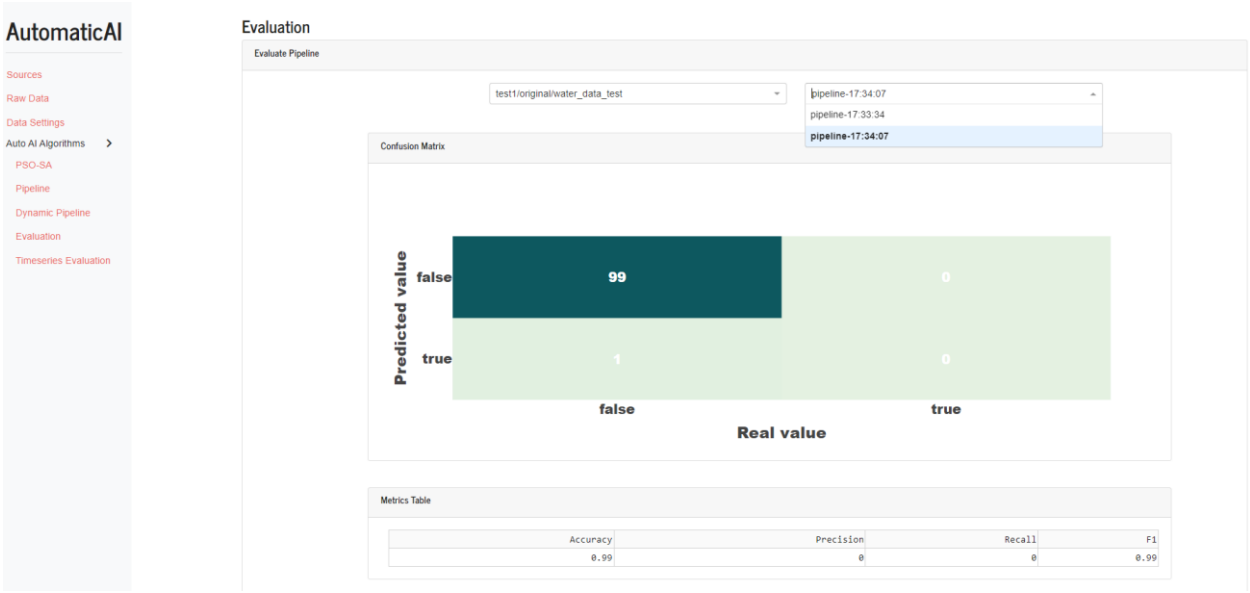


Fig. 71 Evaluarea pipeline-ului de clasificare

În cazul seriilor de timp, rezultatele sunt afișate folosind o diagramă de tip linechart (Fig. 72).

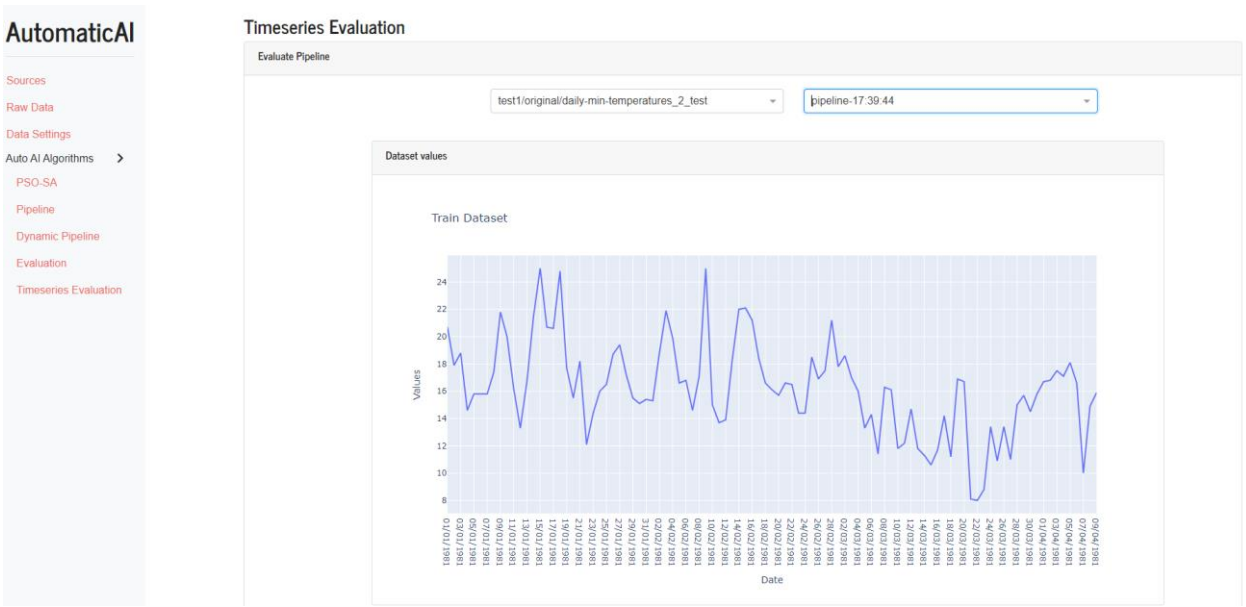


Fig. 72 Evaluarea pipeline-ului de procesare a seriilor de timp

6.2.2 Serviciul de analiză a datelor (UPB.A)

Serviciul de analiză a datelor este compus din următoarele componente la nivel de implementare:

- Implementare Python:
  - Script-uri Python;
  - Date preluate din fișiere CSV;
- Componente:
  - Proiect clusterizare;

  - Proiect clasificare;
- Biblioteci și framework-uri:
  - scikit-learn, Keras, Tensorflow;
  - Matplotlib.

Metode clusterizare

Deoarece datele reprezintă consumul mediu săptămânal pe oră, pe întreaga perioadă de timp, pentru fiecare nod de măsurare, a fost utilizată o metodă de clustering K-Means pentru a identifica tiparele medii săptămânale de consum.

Pentru a identifica numărul optim de clustere pentru setul de date disponibil, am aplicat metoda Elbow pe baza WCSS. Datele colectate de la senzorii IoT au fost stocate în fișiere de tip CSV, pentru a efectua cu ușurință procesarea acestora, cu scopul final de a oferi un rezultat cât mai apropiat de cel real.

Gruparea datelor ne-etichetate este efectuată folosind modulul *sklearn.cluster* din biblioteca scikit-learn. Fiecare algoritm de grupare vine în două variante: o clasă, care implementează metoda *fit* pentru antrenarea modelului și o funcție, care returnează etichetele identificate, corespunzătoare diferitelor clustere, ce pot fi găsite în atributul labels_.

Algoritmul K-Means grupează datele încercând să separe eșantioanele în n grupuri de varianță egală, minimizând WCSS. Acest algoritm necesită specificarea numărului de clustere. Este scalabil pentru un număr mare de date și a fost utilizat în multe domenii diferite.

Metode clasificare

Metodele de clasificare se împart în 2 clase, folosind biblioteci și framework-uri diferite.

Scikit-learn

În prima categorie intră arborii de decizie (DT) și Random Forest, framework-ul folosit pentru acestea fiind scikit-learn.

Arborii de decizie (DT) reprezintă o metodă de învățare supervizată neparametrică utilizată pentru clasificare și regresie. Scopul este de a crea un model care prezice valoarea unei variabile țintă prin învățarea unor reguli simple de decizie deduse din caracteristicile datelor.

Ca implementare, *DecisionTreeClassifier* este utilizat pentru clasificarea multi-clasă pe setul de date etichetat. Ca și în cazul altor clasificatoare, *DecisionTreeClassifier* primește ca intrare două matrici: X, de formă (n_samples, n_features) care conține datele de antrenare și Y, de formă (n_samples,), care conține etichetele clasei pentru datele de antrenare. După antrenarea modelului prin metoda *fit*, acesta este folosit pentru predicția claselor, folosind metoda *predict* [22].

Pentru a îmbunătăți robustețea, se folosesc metode ansamblu, prin care se combină predicțiile mai multor estimatori. O astfel de metodă este reprezentată de Random Forest (*RandomForestClassifier*). Fiecare arbore decizional din ansamblu este construit dintr-un eșantion extras din setul de antrenare. Algoritmul Random Forest realizează o variație redusă prin combinarea diverșilor arbori de decizie, rezultând astfel un model general mai bun. Dimensiunea

[22] https://scikit-learn.org/stable/modules/tree.html

eșantionului este controlată cu parametrul max_samples dacă bootstrap=True (implicit); în caz contrar, întregul set de date este folosit pentru a construi fiecare arbore.

În plus, la împărțirea fiecărui nod în timpul construcției unui arbore, cea mai bună împărțire este găsită fie din toate caracteristicile de intrare, fie dintr-un subset aleatoriu de dimensiune max_features. (Consultați instrucțiunile de reglare a parametrilor pentru mai multe detalii [23]).

TensorFlow. Keras

Ceilalți algoritmi folosiți se regăsesc în framework-ul Keras construit peste TensorFlow. În timp ce TensorFlow este un strat de infrastructură pentru programare diferențiabilă, care se ocupă de tensori, variabile și gradienți, Keras este o interfață cu utilizatorul pentru învățare profundă, care se ocupă de straturi, modele, optimizatori, funcții de pierdere, metrici și multe altele.

Keras servește ca API de nivel înalt pentru TensorFlow. Clasa Layer este abstractizarea fundamentală în Keras. Un Layer încapsulează o stare (ponderi) și unele calcule (definite în metoda de apelare). În primă fază este construit modelul rețelei neuronale:

```
# add input layer
model.add(Dense(hsize, input_dim=input_dim, activation='relu'))

# add dropout to hidden layers to prevent overfitting
model.add(Dropout(0.5))
model.add(Dense(hsize, activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(hsize, activation='relu'))
model.add(Dropout(0.5))

# add output layer
model.add(Dense(output_dim, activation=activation_fn))
```

Apoi modelul este antrenat pe setul de date, distribuit în date de antrenare și de validare:

```
early_stopping_monitor = EarlyStopping(patience=50)

# fit the keras model on the dataset
h = model.fit(X, y, validation_split=0.2, epochs=epochs,
 batch_size=batch_size, verbose=1, callbacks=[early_stopping_monitor])
```

Acuratețea modelului este evaluată pe setul de date de test:

```
_, accuracy = model.evaluate(X, y, batch_size=batch_size, verbose=1)
```

Predicția se realizează pe baza datelor de intrare, care pot fi neetichetate:

[23] https://scikit-learn.org/stable/modules/ensemble.html#random-forest-parameters

```
predictions = model.predict(X)
```

Vizualizare

Matplotlib

Matplotlib este o bibliotecă de plotare pentru limbajul de programare Python, ce oferă un API orientat pe obiecte pentru încorporarea graficelor în aplicații de vizualizare a datelor. Pyplot este un modul Matplotlib care oferă o interfață asemănătoare MATLAB. Matplotlib este conceput pentru a fi la fel de utilizabil ca MATLAB, cu posibilitatea de a folosi Python și avantajul de a fi gratuit și open-source. Matplotlib este folosit în cadrul serviciului de analiză a datelor pentru:

- reprezentarea seriilor de timp;
- vizualizarea datelor procesate în coordonate 2D;
- vizualizarea datelor procesate sub formă de hartă a culorilor:
- analiză comparativă prin grafic cu bare.

Integrare

În urma integrării serviciilor oferite de UPB și UTCN, partea legată de imputarea datelor și scalarea acestora nu va mai fi realizată la nivelul UPB, ci va fi preluată de la serviciul AutomaticAI, care, pe lângă imputarea datelor, se mai ocupă și de detecția anomaliilor și de predicțiile ce trebuie făcute.

6.3 Crowdsensing (CS)

Platforma de crowdsensing este implementată de o aplicație mobilă, bazată pe o platformă de explorare urbană prin gamificare, dezvoltată în colaborare cu LEPLACE GLOBAL, un startup românesc din domeniul aplicațiilor de realitate augmentată. Participanții / cetățenii pot raporta probleme legate de utilitățile de apă și le pot plasa pe hartă. Gamificarea este folosită pentru a crește nivelul de participare și interacțiune cu mediul, în timp ce design-ul fluid și aspectul vizual asigură un mediu de joc atractiv, cu implicații în lumea reală. Pagina principală este prezentată în Fig. 73, cu locația problemelor și a participanților conform configurației experimentale.

Deși există diferite tipuri de probleme care pot fi găsite în infrastructura urbană de apă, vizualizarea hărții dezvăluie probleme din apropiere, așa cum sunt raportate de participanți: inundații, probleme de alimentare cu apă (e.g. scurgeri sau întreruperi ale serviciului), poluare și alte pericole. Locațiile participanților sunt afișate pe hartă în scop experimental, în timp ce în aplicațiile din lumea reală, colectarea datelor de locație în timp real poate fi interzisă de legile de confidențialitate, conform cazului de utilizare.

Modelul crowdsensing prezentat în etapa de proiectare poate fi utilizat pentru a defini alocarea sarcinilor pentru fiecare participant și pentru a oferi mecanismul de stimulare. Problemele raportate pot fi înregistrate în blockchain, pentru un nivel sporit de încredere pentru părțile interesate, prin stocarea securizată a datelor și transparență.

Fig. 73 Aplicația Watergame pentru raportarea incidentelor

Componentele propuse abordează problemele existente în aplicațiile mobile de crowdsensing, precum problemele legate de motivație, încredere și acoperire, din trei perspective diferite. Prin urmare, utilitatea globală a platformei, adică acoperirea și costul asociat cu raportările, este influențată de: (1) modelul crowdsensing; (2) un design de joc atractiv pentru a crește nivelul de participare; (3) nivelul de încredere perceput, asigurat de tehnologia blockchain.

Pentru a asigura performanța și scalabilitatea soluției, server-ul aplicației este găzduit în cloud (AWS) și au fost proiectate o serie de componente și optimizări de arhitectură precum integrarea unei memorii cache REDIS, optimizări la nivel de bază de date (indexarea tabelelor, refactorizarea bazei de date în funcție de cazurile de utilizare), optimizări în transferul de date (stocare temporară, compactarea structurilor de date arborescente). Aceste optimizări precum și integrarea cu platforma de explorare urbană vor fi prezentate mai pe larg în etapa de integrare.

6.4 Blockchain (B)

6.4.1 Aplicatie Web – Oferta de moneda inițială

Aplicația web dezvoltată permite clienților achiziționarea de criptomonede înainte de lansarea soluției, așa cum este prezentat în Fig. 74. Monedele achiziționate vor putea fi folosite pentru încheierea de contracte cu beneficii suplimentare sau ca investiție financiară, datorită faptului că în timp valoarea monedelor se va aprecia, ulterior acestea putând fi vândute, prin intermediul soluțiilor de exchange ce vor adopta moneda WGT.

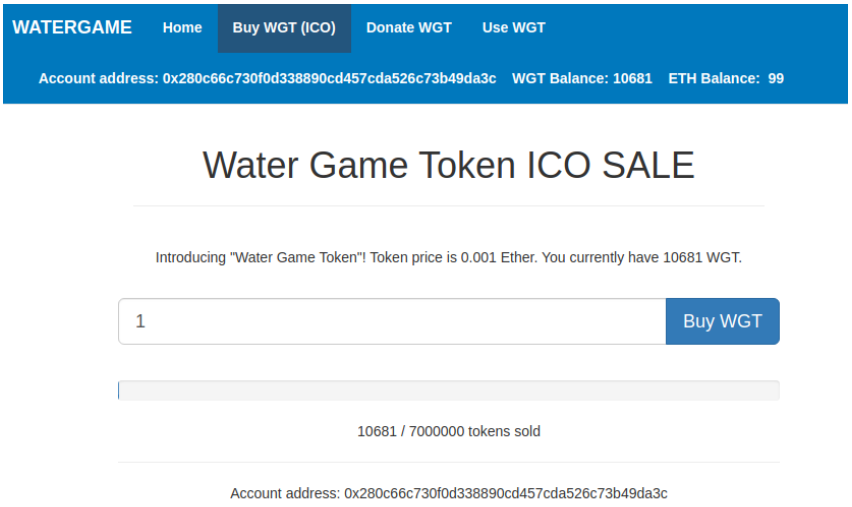


Fig. 74 Interfața sistemului blockchain. ICO

6.4.2 Aplicație Web – Transfer criptomonedă

Aplicația web dezvoltată permite clienților sa transfere o cantitate de criptomonedă din portofelul electronic propriu la adresa unui prieten așa cum este prezentat în Fig. 75. Astfel soluția va crea o comunitate de utilizatori ce interacționează între ei prin token.

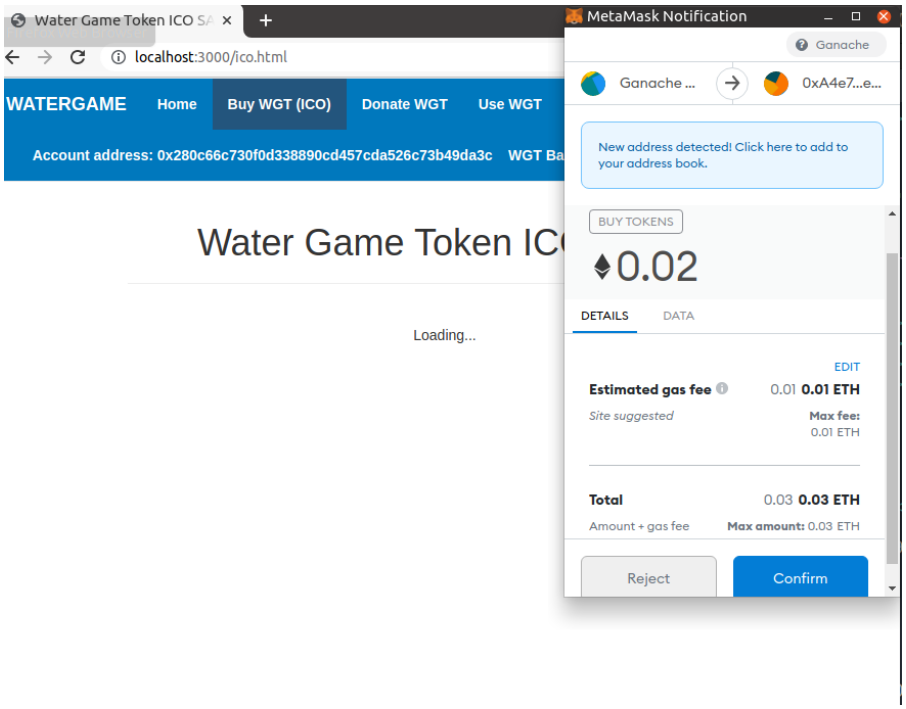


Fig. 75 Interfața sistemului Blockchain. Integrare cu Ethereum

În Fig. 75 este prezentat modul în care se realizează o tranzacție cu criptomoneda WGT folosind MetaMask, împreună cu mesajul de succes aferent acestei tranzacții (Fig. 76).

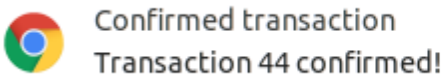


Fig. 76 Interfața sistemului Blockchain. Confirmarea tranzacției.

6.4.3 Funcționalitate REST

Pentru integrarea tuturor componentelor sistemului cu zona de blockchain este necesară implementarea unei modalități de comunicare bazate pe un REST API prin care se vor putea declanșa diverse acțiuni / tranzacții în funcție de output-ul celorlalte sisteme (de ex.: acordarea de recompense).

Astfel, în următoarea etapă, această componentă va permite inițializarea unei tranzacții prin apeluri REST care să transfere o cantitate de criptomonede din portofelul electronic al furnizorului (sau administratorului) către un utilizator drept recompensă (de ex.: realizarea unui consum optim pentru o lună sau sesizarea unui incident valid în aplicația de crowdsensing).

7 Pregătirea datelor

În scenariile prezentate anterior colectarea datelor presupune obținerea acestora de la sursele externe de date. Astfel, datele primare sunt furnizate de la debitmetre conectate într-unul sau mai multe puncte din cadrul aceleiași locații.

7.1.1 Metode de procesare în edge (UTCN.C)

La nivelul nodurilor din cadrul locațiilor de măsurare se filtrează valorile din afara intervalului de măsurare suportat de debitmetrele instalate în cadrul locațiilor. Aceste filtrări ale valorilor se fac în baza fișei tehnice a debitmetrelor excluzând orice citiri care nu se află în intervalul de referință specificat de producătorii debitmetrelor. Astfel se reduce latența și utilizarea lățimii de bandă pentru transmiterea datelor procesate către serviciile de colectare și stocare din cloud. Transmiterea datelor sub forma brută (raw) spre serviciile de colectare a datelor din cloud este lentă și nu profită de posibilitatea utilizării puterii de procesare locale din cadrul nodurilor. Profitând de avantajele procesării în edge se face analiza în timp real a datelor, reducerea latenței și posibilitatea de a acționa în timp real pe baza datelor procesate.

O informație extrem de importantă ca urmare a monitorizării consumului de apă constă în detecția situațiilor de scurgeri, de conducte sparte, de pierderi în rețeaua de distribuție aflată înaintea punctelor de citire a debitelor și acționarea în consecință pentru limitarea sau stoparea pierderilor de apă.

În locațiile unde sunt montate mai multe debitmetre la diferiți consumatori din cadrul locației cât și pe conducta principală de alimentare cu apă a locației, la nivelul rețelei de debitmetre conectate în unul sau mai multe noduri se pot determina cu ușurință pierderile din interiorul locației și se poate interveni cu ajutorul dispozitivelor de tip electrovalve sau electroventile pentru oprirea alimentării cu apă.

7.1.2 Metode de preprocesare premergătoare analizei datelor (UTCN.C)

Majoritatea seturilor de date din lumea reală sunt foarte susceptibile la date lipsă, date inconsistente și zgomotoase din cauza originii lor eterogene. Aplicarea algoritmilor de învățare automată pe aceste date zgomotoase ar da rezultate de calitate scăzută, deoarece nu ar putea identifica tiparele în mod eficient. Preprocesarea datelor este, prin urmare, importantă pentru a îmbunătăți calitatea generală a datelor, astfel îmbunătățind rezultate obținute prin algoritmi de învățare automată.

În acest proiect am analizat și am implementat metode / algoritmi pentru următoarele categorii mari de metode de preprocesare: Curățarea datelor; Transformarea datelor și Echilibrarea datelor.

Curățarea datelor

Scopul curățării datelor este de a oferi seturi de date simple, complete și clare pentru învățarea automată.

Date lipsă

De multe ori trebuie să lucrăm cu date de intrare în care lipsesc anumite date, ceea ce poate strica performanța modelului de învățare automată. Pentru a rezolva această problemă avem următoarele posibilități incluse în platforma noastră de procesare și analiză a datelor:

- Ștergerea rândurilor cu valori lipsă;
- Completarea valorilor lipsă folosind media, mediana sau valori implicite;
- Utilizarea algoritmilor care acceptă valorile lipsă;
- Predicția valorilor lipsă.

Transformarea datelor

Normalizare / Scalare

Scopul normalizării este de a schimba valorile coloanelor numerice din setul de date la o scară comună, fără a distorsiona diferențele în intervalele de valori. Aici putem vorbi de metode care setează media datelor la 0 și deviația standard la 1 și metode care scalează intervalul de date la [0,1], metode care micșorează / extind un vector (un rând de date poate fi văzut ca un vector D-dimensional) la o sferă unitară.

Echilibrarea datelor

Algoritmii folosiți în acest proiect pentru rebalansarea / reechilibrarea setului de date sunt SMOTE (Chawla et al., 2002) pentru a supra-eșantiona clasa minoritară (utilizând valori sintetice, generate) și RUID (Prusa et al., 2015) pentru sub-eșantionerea clasei majoritare, încercând astfel să reechilibrăm setul de date.

7.1.3 Metode de stocare și vizualizare (UPB)

Datele colectate prin MQTT sunt înregistrate în MySQL pentru fiecare senzor în parte. Schema bazei de date este prezentată în Fig. 77 și presupune stocarea datelor de consum în tabelul *sensor_data* și a informațiilor despre senzori în tabelul *sensor*. Datele colectate sunt identificate prin ID-ul senzorului (i.e. modul de achiziție) și canalul măsurat (i.e. poziția debitmetrului atașat modulului de achiziție). Aceste informații sunt preluate din mesajul trimis de senzor prin MQTT, iar referința de timp (*timestamp*) este introdusă de server-ul aplicației.

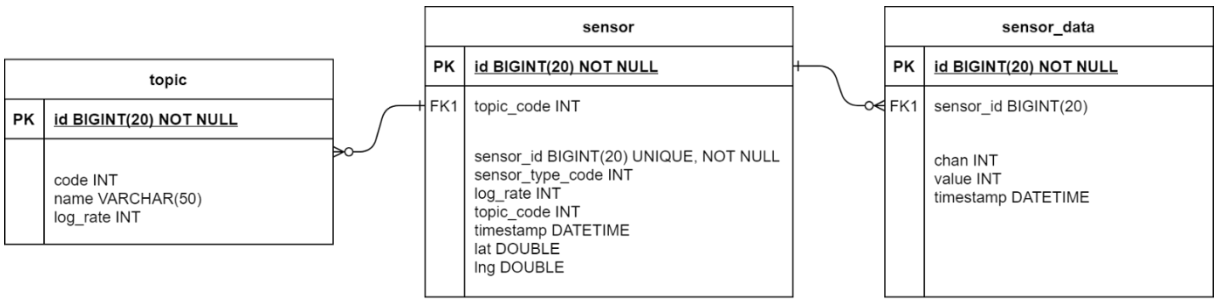


Fig. 77 Schema bazei de date

Atunci când sunt solicitate pentru vizualizare, datele sunt pregătite și grupate de server-ul aplicației pentru a fi aduse în format CSV sau JSON. Se poate astfel interoga baza de date și pot fi incluse filtre precum ID-ul senzorului, canalul de măsură, intervalul de timp sau numărul de

înregistrări solicitate. În Fig. 78 este prezentat modul de testare a endpoint-urilor, din interfața Postman[24].

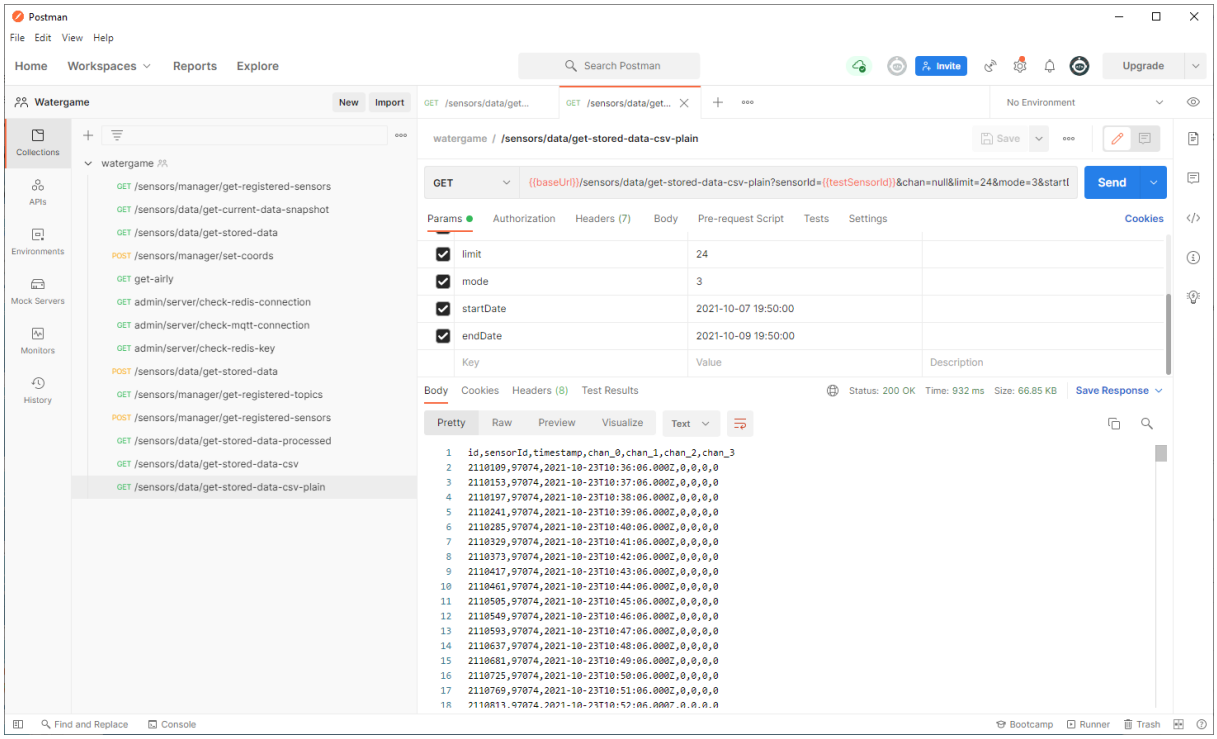


Fig. 78 Vizualizarea datelor în interfața Postman

Datele pot fi vizualizate în aplicația WaterMonitoringSystem sau prelucrate conform scenariilor DSS în Python. Prin interogarea bazei de date și prelucrarea primară pe server, se obțin datele în format CSV care sunt apoi procesate în Python.

În Fig. 79 sunt prezentate datele colectate de la un modul de achiziție instalat într-o baie pe o perioadă de 2 săptămâni, care măsoară volumul de apă consumat pe unitatea de timp pentru senzorii instalați: apă rece, apă caldă, toaletă, duș. În mod similar, datele de consum instantaneu (debit) sunt prezentate în Fig. 80 pentru același modul instalat.

[24] https://www.postman.com/

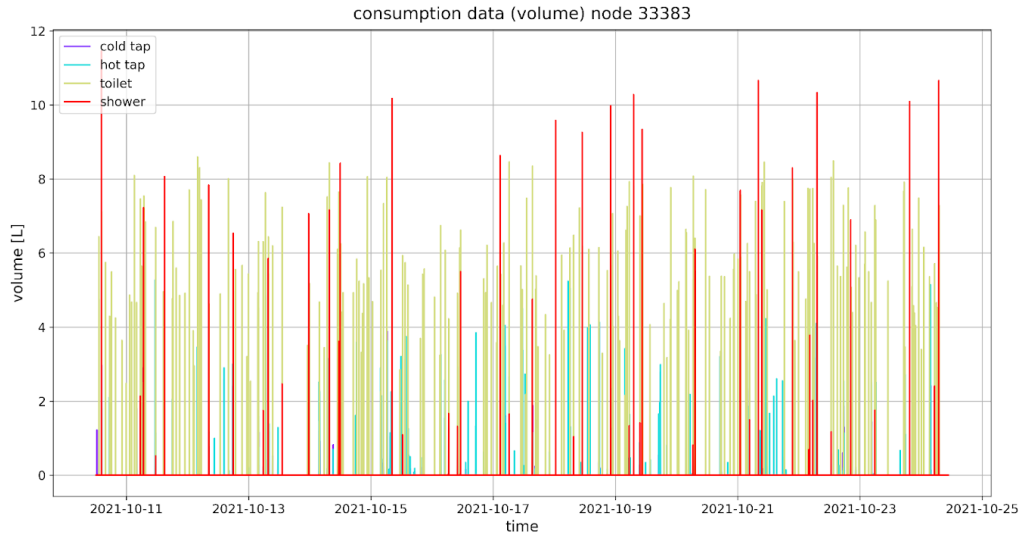


Fig. 79 Exemplu date consum (volum / consum total pe eșantion)

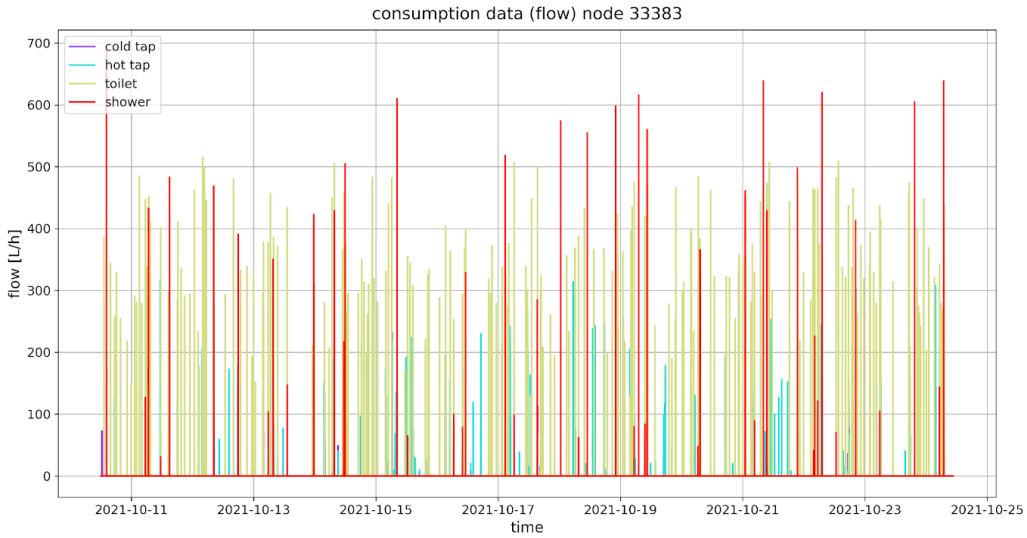


Fig. 80 Exemplu date consum (debit / consum instantaneu)

Pentru identificarea tiparelor de consum pe o perioadă mai lungă de timp, a fost aplicat un filtru de mediere, rezultatele fiind prezentate în Fig. 81. În timp ce consumul de apă la chiuvetă este relativ constant iar consumul de apă la duș este similar ca valoare medie, se observă un consum de 3-4 ori mai mare în alimentarea toaletei. În acest caz, sistemul DSS poate furniza indicii pentru a contribui la reducerea consumului de apă, cum ar fi utilizarea unui bazin mai mic la toaletă.

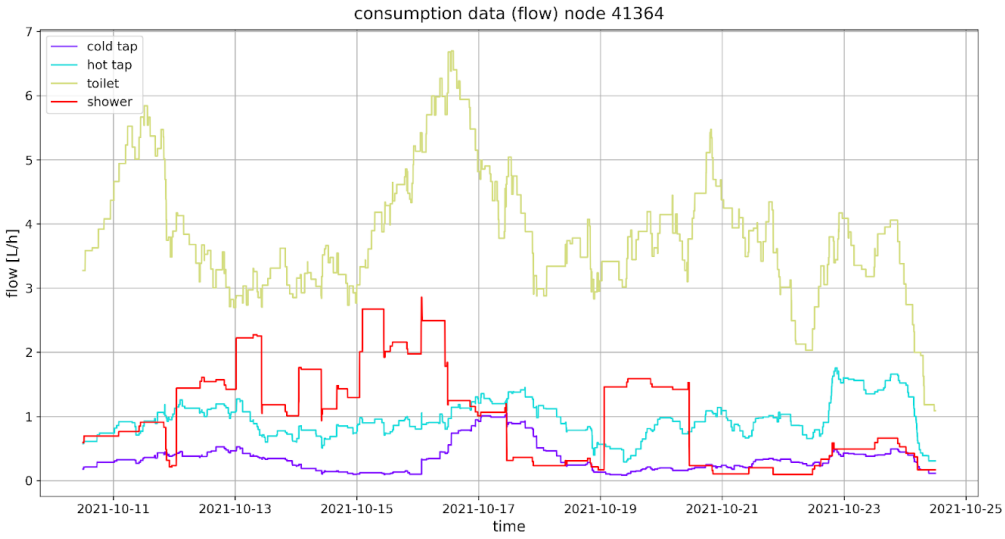


Fig. 81 Exemplu date consum (debit / consum instantaneu, filtru de mediere)

8 Concluzii

Gestionarea resurselor de apă în sistemele de distribuție este un aspect important al durabilității în zonele metropolitane. Prin utilizarea IWM-urilor și a metodelor de învățare nesupervizate, cererea consumatorilor poate fi luată în considerare cu precizie la proiectarea unui DSS la scară largă. Într-o rețea inteligentă de apă, obiceiurile, nevoile și comportamentele consumatorilor ar trebui luate în considerare pentru o gestionare optimă a resurselor, în timp ce consumatorii pot avea ocazia să monitorizeze și să ia decizii cu privire la conservarea resurselor de apă. Diferite clase de tipare de consumatori au fost extrase din datele săptămânale ale cererii, subliniind tendințele, variația și profilurile generale. În acest sens, modelele identificate pot oferi o mai bună înțelegere a comportamentului consumatorului în direcția dezvoltării de soluții stimulative și centrate pe consumatori pentru optimizarea resurselor de apă.

Din punctul de vedere al furnizorilor, acest studiu îi susține în procesul de luare a deciziilor, analizând datele extrase și profilurile consumatorilor în funcție de diverși factori, cum ar fi locația geo-spațială, în timp ce standardul de viață și condițiile socio-economice pot fi, de asemenea, luate în considerare. Pe baza datelor, statisticile și rapoartele pot oferi o imagine de ansamblu asupra consumului de apă, în timp ce poate fi definit un sistem de suport al deciziilor pentru a ajuta la conservarea resurselor de apă. Localizarea geografică a consumatorilor oferă, de asemenea, o imagine de ansamblu asupra rețelei de distribuție a apei în ceea ce privește identificarea posibilelor strategii combinate pentru grupurile de consumatori.

Soluțiile propuse evidențiază regiunile în care consumul de apă depășește limitele dinamice ale consumului optim identificat în rețea, în timp ce consumatorii pot primi feedback și recomandări pe baza rezultatelor. Mai mult, realizarea unui model de consum echilibrat într-o regiune geografică poate îmbunătăți sustenabilitatea infrastructurii de apă, luând în considerare efectul combinat al modelelor de consumatori, reducând astfel nevoia de recomandări excesive pentru consumatorii individuali.

În etapa curentă a proiectului s-au identificat cazurile de utilizare, s-a proiectat arhitectura generală a sistemului, s-au proiectat și implementat componentele sistemului, urmând ca în etapa următoare să se integreze componentele dezvoltate de cele două universități (UPB și UTCN) și să se realizeze testarea sistemului având componentele integrate.

Bibliografie

Alzen, J. L., Langdon, L. S., & Otero, V. K. (2018). A logistic regression investigation of the relationship between the Learning Assistant model and failure rates in introductory STEM courses. International journal of STEM education, 5(1), 1-12, doi 10.1186/s40594-018-0152-1.

Amazon Web Services. (2021). Overview of Amazon Web Services AWS Whitepaper. https://docs.aws.amazon.com/whitepapers/latest/aws-overview/aws-overview.pdf

Arsene, D., Predescu, A., Truică, C.-O., Apostol, E.-S., Mocanu, M., & Chiru, C. (2021). Profiling consumers in a water distribution network using K-Means clustering and multiple pre-processing methods. 2021 13th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), 1–6. https://doi.org/10.1109/ECAI52376.2021.9515193

Austin Water. (2021), Austin Water - Residential Water Consumption, https://data.austintexas.gov/Utilitiesand-City-Services/Austin-Water-Residential-WaterConsumption/sxk7-7k6z

Bell, C. (2016). MySQL for the Internet of Things. https://doi.org/10.1007/978-1-4842-1293-6

Bento, A. (2018). IoT: NodeMCU 12e X Arduino Uno, Results of an experimental and comparative survey. 6, 45–56.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32, DOI: 10.1023/A:1010933404324.

Chandrasekaran, S., Friese, M., Stork, J., Rebolledo, M. & Bartz-Beielstein, T. (2017) Gecco Challange 2017, https://www.spotseven.de/gecco/gecco-challenge/gecco-challenge-2017/

Chatzigeorgakidis, G. (2018). Household Water Consumption (Average), HELIX, 2018. https://data.hellenicdataservice.gr/dataset/236a0883-91b6-4f57-af9a-8d54ae7b154e (accessed Apr. 20, 2021).

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357, DOI: 10.1613/jair.953.

Chen, S., Tang, X., Wang, H., Zhao, H., & Guo, M. (2016). Towards Scalable and Reliable In-Memory Storage System: A Case Study with Redis. https://doi.org/10.1109/TrustCom.2016.0255

ConsenSys Software Inc. (2021). Truffle | Overview | Documentation | Truffle Suite. https://www.trufflesuite.com/docs/truffle/overview

Cunningham, P., & Delany, S. J. (2007). k-Nearest neighbour classifiers. Mult Classif Syst., 54, DOI: 10.1145/3459665.

Czako, Z., Sebestyen, G., & Hangan, A. (2021). AutomaticAI-A Hybrid Approach for Automatic Artificial Intelligence Algorithm Selection and Hyperparameter Tuning. Expert Systems with Applications, 115225, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2021.115225.

Evgeniou, T., & Pontil, M. (1999, July). Support vector machines: Theory and applications. In Advanced Course on Artificial Intelligence (pp. 249-257). Springer, Berlin, Heidelberg, DOI: 10.1007/3-540-44673-7_12.

ethereum.org. (2021). Token ERC-20 Standard. https://ethereum.org/ro/developers/docs/standards/tokens/erc-20/

Ethereum Revision a57dd3c5. (2016). Getting Started - web3.js 1.0.0 documentation. https://web3js.readthedocs.io/en/v1.5.2/getting-started.html

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine learning, 63(1), 3-42, DOI: 10.1007/s10994-006-6226-1.

Google. (2021a). Firebase Realtime Database. https://firebase.google.com/docs/database

Google. (2021b). Vehicle Routing Problem | OR-Tools | Google Developers. https://developers.google.com/optimization/routing/vrp

Hanum, F., Hartono, A. P., & Bakhtiar, T. (2018). On the multiple depots vehicle routing problem with heterogeneous fleet capacity and velocity. IOP Conference Series: Materials Science and Engineering, 332, 12052. https://doi.org/10.1088/1757-899x/332/1/012052

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

IBM. (2021). Getting to know MQTT. https://developer.ibm.com/articles/iot-mqtt-why-good-for-iot/

Karray, F., Triki, M., Jmal, W., Abid, M., & Obeid, A. (2018). WiRoTip: an IoT-based Wireless Sensor Network for Water Pipeline Monitoring. International Journal of Electrical and Computer Engineering (IJECE), 8, 3250. https://doi.org/10.11591/ijece.v8i5.pp3250-3258

Khasawneh, M., Kajman, I., Alkhudaidy, R., & Althubyani, A. (2014). A Survey on Wi-Fi Protocols: WPA and WPA2 (Vol. 420). https://doi.org/10.1007/978-3-642-54525-2_44

Khawas, C., & Shah, P. (2018). Application of Firebase in Android App Development-A Study. International Journal of Computer Applications, 179, 49–53. https://doi.org/10.5120/ijca2018917200

Lu, W., Li, J., Li, Y., Sun, A., & Wang, J. (2020). A CNN-LSTM-based model to forecast stock prices. Complexity, 2020.

Makowski, L., & Ostroy, J. M. (1987). Vickrey-Clarke-Groves mechanisms and perfect competition. Journal of Economic Theory, 42(2), 244–261. https://doi.org/10.1016/0022-0531(87)90087-1

Masse, M. (2011). REST API Design Rulebook: Designing Consistent RESTful Web Service Interfaces. " O'Reilly Media, Inc."

MetaMask. (2021). MetaMask - A crypto wallet & gateway to blockchain apps. https://metamask.io/index.html

Mishra, B., & Kertesz, A. (2020). The Use of MQTT in M2M and IoT Systems: A Survey. IEEE Access, 8, 201071–201086. https://doi.org/10.1109/ACCESS.2020.3035849

Moroney, L. (2017). The Firebase Realtime Database (pp. 51–71). https://doi.org/10.1007/978-1-4842-2943-9_3

Murtagh, F. (1991). Multilayer perceptrons for classification and regression. Neurocomputing, 2(5-6), 183-197, ISSN 0925-2312, https://doi.org/10.1016/0925-2312(91)90023-5.

Neto, A. (2020). Developing for Ethereum, Test networks, Standalone Nodes, Solidity, and the Remix IDE. Medium. https://medium.com/@arlindont/developing-for-ethereum-9a0551f2b9d9

NodeMcu Team. (2018). NodeMcu -- An open-source firmware based on ESP8266 wifi-soc. https://www.nodemcu.com/index_en.html

Parihar, Y. S. (2019). Internet of Things and Nodemcu A review of use of Nodemcu ESP8266 in IoT products. 6, 1085.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Portet, S. (2020). A primer on model selection using the Akaike Information Criterion. Infectious Disease Modelling, 5, 111-128, doi 10.1016/j.idm.2019.12.010.

Predescu, A., & Mocanu, M. (2019). Increasing Collaboration and Participation Through Serious Gaming for Improving the Quality of Service in Urban Water Infrastructure. In Lecture Notes in Business Information Processing: Vol. 373 LNBIP. https://doi.org/10.1007/978-3-030-36691-9_49

Predescu, A., Arsene, D., Pahonțu, B., Mocanu, M., & Chiru, C. (2021). A Serious Gaming Approach for Crowdsensing in Urban Water Infrastructure with Blockchain Support. Applied Sciences, 11(4). https://doi.org/10.3390/app11041449

Prusa, J., Khoshgoftaar, T. M., Dittman, D. J., & Napolitano, A. (2015, August). Using random undersampling to alleviate class imbalance on tweet sentiment data. In 2015 IEEE international conference on information reuse and integration (pp. 197-202). DOI: 10.1109/IRI.2015.39.

scikit-learn developers (BSD License). (n.d.). scikit-learn. http://scikit-learn.org/stable/

Sessa, P. G., Walton, N., & Kamgarpour, M. (2017). Exploring the Vickrey-Clarke-Groves Mechanism for Electricity Markets. IFAC-PapersOnLine, 50(1), 189–194. https://doi.org/10.1016/j.ifacol.2017.08.032

Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 404, 132306, DOI: 10.1016/j.physd.2019.132306.

Tao, X., & Song, W. (2018). Efficient Path Planning and Truthful Incentive Mechanism Design for Mobile Crowdsensing. Sensors (Basel, Switzerland), 18(12), 4408. https://doi.org/10.3390/s18124408

Topîrceanu, A., & Grosseck, G. (2017). Decision tree learning used for the classification of student archetypes in online courses. Procedia Computer Science, 112, 51-60, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2017.08.021.

Wingerath, W. (2017). A Real-Time Database Survey: The Architecture of Meteor, RethinkDB, Parse & Firebase (Medium (ed.)).

Xingjian, S. H. I., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In Advances in neural information processing systems (pp. 802-810).

Yermal, L., & Balasubramanian, P. (2017, December). Application of auto arima model for forecasting returns on minute wise amalgamated data in nse. In 2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC) (pp. 1-5). IEEE