

IoT adatgyűjtési és elemzési platform a vízellátó infrastruktúra-rendszerek anomáliáinak észlelésére

IoT data collection and analysis platform used for anomaly detection in water supply infrastructure systems

CZAKÓ Zoltán

zoltan.czako@cs.utcluj.ro

Számítástechnikai Tanszék, Kolozsvári Műszaki Egyetem, Kolozsvár, Románia

Vezető tanárok: SEBESTYEN Gheorghe, HANGAN Anca

Kivonat

A vízellátás kritikus közszolgáltatás, amely közvetlen hatással van a lakosság jólétére és biztonságára. A modern vízellátási infrastruktúrák összetett kiberfizikai rendszerek, amelyek fizikai eszközöket (pl. csövek, szivattyúk, tartályok, tisztítóállomások stb.), valamint automatizálási és felügyeleti eszközöket foglalnak magukba. Az ilyen rendszerek fő kihívásai: a minőségi víz folyamatos ellátása, a vízkészletek optimális elosztása és a meglévő erőforrások hatékony felhasználása. Az anomáliák észlelése a vízellátó infrastruktúrákban fontos szempont, amely befolyásolja a teljes rendszer megfelelő működését és biztonságát. Az ilyen infrastruktúrák egyre összetettebbé válásával az anomáliák észlelését automatizált módon kell végrehajtani. Ebben a cikkben egy olyan IoT infrastruktúrát és platformot mutatunk be, amely lehetővé teszi az adatok begyűjtését különböző szenzoroktól és azok automatikus elemzését elősegíti különböző mesterséges intelligencia algoritmusok felhasználásával. Ezen platform segítségével létrehoztunk egy olyan mesterséges intelligencia modellt, amely képes automatikusan detektálni a víz minőségének különböző változásait, melyeket anomáliákként könyvelhetünk el.

Kulcsszavak: IoT, anomáliák, mesterséges intelligencia, vízminőség

1. BEVEZETŐ

Az Internet of Things (IoT) egy bevett irányzat az internetes kommunikáció területén, amelynek célja a szükséges keretrendszerek és technológiák kifejlesztése különböző eszközök internethez való csatlakoztatásához, a szenzoroktól és végrehajtó elemektől kezdve, háztartási gépeken és épületautomatizáláson át az ipari berendezésekig. Az IoT-eszközök már egy ideje hatalmas mennyiségű adatot generálnak, amelyeket a rendszerek, közösségek és egyének viselkedésével kapcsolatos információk kinyerésére, előrejelzések készítésére, döntéshozatal támogatására vagy anomáliák észlelésére használnak. Ekkora mennyiségű adatnál a legfőbb kihívást az eszközazonosítás, adatátvitel, valamint az adatok tárolása és főleg az adatok elemzése jelenti. Nagy mennyiségű tárolókapacitásra, valamint speciális adatmodellekre van szükség, amelyek képesek kezelni ezt a nagy mennyiségű adatot. Nem elég az adatokat feldolgozás és elemzés nélkül tárolni, hiszen elemzés nélkül nincs hozzáadott értékük, elemzés nélkül az adatok értéktelenek. De a nagy mennyiségű nyers adatból hasznos információk kinyerése és elemzése előfeldolgozást és intelligens elemzési algoritmusokat igényel, amelyek időigényesek és sok erőforrást igényelnek.

A víz jelentős erőforrás egy nemzet számára, és kezelése befolyásolhatja egy régió társadalmi-gazdasági fejlődését. A múlt században a földkerekség számos régióját drasztikusan befolyásolta a vízhiány vagy az ivóvíz rossz minősége. A globális felmelegedés súlyosbítja a vízzel kapcsolatos problémákat, és a jövőben ezek a problémák kritikussá válnak. A víz körforgásának minden szakaszában, a begyűjtéstől és kezeléstől az elosztásig és újrahasznosításig komplex kiberfizikai rendszereknek kell biztosítaniuk az egyensúlyt a kereslet és a kínálat között, és garantálniuk kell a nemzeti és nemzetközi szabványoknak megfelelő minőségi paramétereket. Ezek a funkciók megkövetelik a víz fizikai és kémiai paramétereinek monitorozását (különböző szakaszokban), valós idejű módon, valamint manuális vagy automatizált reakciót a rendszer egyensúlyi állapotában bekövetkezett bármilyen változás következtében.

Nagy kihívást jelent a normálisnak tekintett állapottól való bármely eltérés automatizált azonosítása, más szóval az anomáliák észlelése. A rendellenes viselkedés több tényező eredménye lehet: a vízforrások megváltozása (pl. áramlás, minőség), az előkészítési vagy elosztási infrastruktúra hibái (pl. hibás érzékelők vagy aktuátorok, csőszivárgások) vagy rosszindulatú támadások. A vízrendszer anomáliájának hatásai a

lakosság jelentős részének jólétét és egészségbiztonságát befolyásolhatják. Ezért minden incidenst (meghibásodást) a lehető legrövidebb időn belül észlelni kell, és lehetőség szerint automatizált módon helyreállítási intézkedéseket kell indítani.

Az anomáliák észlelése a vízrendszerekben, valamint bármely más kiberfizikai rendszerben nem triviális feladat. Ha emberi közreműködéssel valósul meg, akkor sok éves tapasztalatra, technikai készségekre és néha intuícóra van szükség. Az automatizált rendszerekben az emberi anomáliák azonosítási képességét szabályokkal vagy mesterséges intelligencia modellekkel helyettesíthetik, amelyek képesek azonosítani és megtanulni a szükséges szabályokat. A második megközelítést összetett viselkedésű, többváltozós rendszerek esetén alkalmazzuk. Ebben az esetben nagyon nehéz lenne az egész komplex rendszert szigorúan, szabályok szerint leírni, ezért ember számára az anomáliák detektálása szinte lehetetlenné válik.

Célunk, hogy IoT-szolgáltatásokat építsünk ki adatgyűjtéshez, kezeléshez és elemzéshez a cloud infrastruktúra felhasználásával, hogy támogatást nyújtsunk olyan kutatási projektekhez, amelyek IoT-eszközökről gyűjtenek adatokat, és információkat nyernek ki az anomáliák észleléséhez, előrejelzéséhez vagy döntéshozatali támogatásához. Az IoT szolgáltatás komponenseit úgy terveztük, hogy önállóan teljes megoldásként használhatóak legyenek, vagy nyitott, interoperábilis interfészeket keresztül integrálhatóak más eszközökkel. Platformunkat felhasználtuk egy mesterséges intelligencia modell létrehozására, amely képes automatikusan detektálni anomáliákat az ivóvíz minőségi adatainak felhasználásával.

2. KAPCSOLÓDÓ MUNKÁK

A vízrendszerek anomáliáinak észlelésével számos tanulmány foglalkozott, melyekben a szerzők különböző célokat és megközelítéseket javasoltak. Célként említhetjük a vízminőség-ellenőrzést [1][2][3], az infrastruktúra integritását, a megszerzett adatok helyességét vagy a rosszindulatú támadások észlelését [4][5]. Megoldásként megkülönböztethetünk idősorokhoz kapcsolódó módszereket (pl. regressziós és predikciós módszerek), illetve osztályozáson alapuló módszereket. Az utolsó kategóriában számos mesterséges intelligencia módszert javasoltak a vízrendszer normál és deviáns viselkedésének modellezésére. A gépi tanulási technikákat olyan modell betanítására használják, amely képes azonosítani az anomáliákat az összegyűjtött adatkészletben, off-line módon vagy valós időben.

Számos általános osztályozási módszert sikeresen teszteltek kifejezetten a vízrendszerek anomáliáinak kimutatására. Az [1] cikk szerzői különböző osztályozási módszereket alkalmaznak, mint például döntési fákat, logisztikai modellfákat, véletlenszerű erdőket és Hoeffding fákat, hogy osztályozzák a víz minőségét néhány mért paraméter, például lúgosság, pH-szint és vezetőképesség alapján. A J48 elnevezésű módszer, egy prediktív öntanuló döntési fa (a Weka csomagból kölcsönözve) bizonyult a legnagyobb pontosságúnak, körülbelül 94%-os pontosságot elérve.

A [6] cikkben a szerzők számos gépi tanulási modellt alkalmaztak, például logisztikus regressziót, lineáris diszkriminancia analízist, SVM-et, ANN-t, DNN-t (mély neurális hálózat), RNN-t (ismétlődő neurális hálózat) és LSTM-et (hosszú rövid távú memória), a vízminőségi anomáliák azonosítására egy valós városi vízinfrastuktúrából származó adatkészletben. A szerzők a gépi tanulási technikák mellett bemutatják a bemeneti adatok szűréséhez és előfeldolgozásához szükséges algoritmusokat és megoldásokat a zaj, a hiányzó minták és a valós adatokra jellemző egyéb problémák kiküszöbölésére. Ebben a kísérletben az SVM-modell teljesített a legjobban, a legtöbb valódi pozitívat azonosította, hamis pozitívumot nem generált, és az F1 metrika 0,989 értéket érte el. A fő probléma ebben az esetben az volt, hogy a mért vízminőségi paraméterek (redox, klór, pH, vezetőképesség, zavarosság stb.) önmagukban nem utaltak vízminőségi anomáliára, csak ezek meghatározott kombinációi voltak felismerhetők minőségi incidensek esetén.

A vízminőségi adatok anomáliáinak kimutatására szolgáló másik megközelítést a [7] cikk ismerteti. Itt a szerzők egy kettős időben mozgó ablak módszert javasolnak bizonyos minták valós időben történő felismerésére. A módszer a statisztikai modellek és a lineáris autoregresszió kombinációja. Ebben az esetben az adatokat egy valódi folyó megfigyelőállomásról gyűjtötték. A vízminőségi incidensek észlelése szintén a [7] cikk célja volt. A szerzők 3 modellt, az SVM-et és az ANN két változatát hasonlítják össze, és arra a következtetésre jutottak, hogy az SVM összehasonlítható eredményekkel rendelkezik az ANN módszerekkel, de kevesebb időigényrel.

Mindezeket az eredményeket elemezve megállapítható, hogy a vízi infrastruktúrák anomáliái nagy pontossággal kimutathatók SVM, ANN, döntési fákat, ensemble módszerek, rajoptimalizáló és regressziós modellek vagy ezek kombinációi segítségével. A jó eredmények eléréséhez ezeket a modelleket valós infrastruktúrákból gyűjtött adatokkal kell képezni. Az anomáliák nagyon csekély száma a valós adatokban azonban problémákat okozhat a képzési folyamatban. A korábban bemutatott példákban található legjobb

detektálási algoritmusok nagymértékben függenek a mért paraméterektől, az anomáliák gyakoriságától és a bemeneti adatok minőségétől.

Ebben a cikkben egy automatikus módszert javasolunk, amely kiválasztja a legjobban teljesítő algoritmust egy adott adatkészlethez. Ez a módszer úgy hangolja a talált algoritmusok hiperparamétereit, hogy maximalizálja az adott valós eset szempontjából fontosnak tartott metrikákat.

Terveztünk egy algoritmust, amely automatikusan kiválasztja a legjobb modellt és behangolja hiperparamétereit emberi beavatkozás nélkül, és az anomália-észlelő platform kiegészítő elemeként implementáltuk. Ez az algoritmus a részecskeraj-optimalizálást kombinálja a szimulált hűtéssel, hogy kizárólag a bemeneti adatok alapján keresse meg a legjobb modellt. Az anomáliaészlelési platform (melynek neve AutomaticAI [14]) összetevőjének architektúrája mikroszolgáltatásokként épül fel, és egy felhő-infrastruktúrán van telepítve.

3. A PLATFORM ARCHITEKTÚRÁJA

A cloudon keresztül közzétett IoT-szolgáltatásaink integrálja az IoT-eszközökről származó adatok gyűjtésére, tárolására és elemzésére szolgáló funkciókat. Több összetevőt tartalmaz:

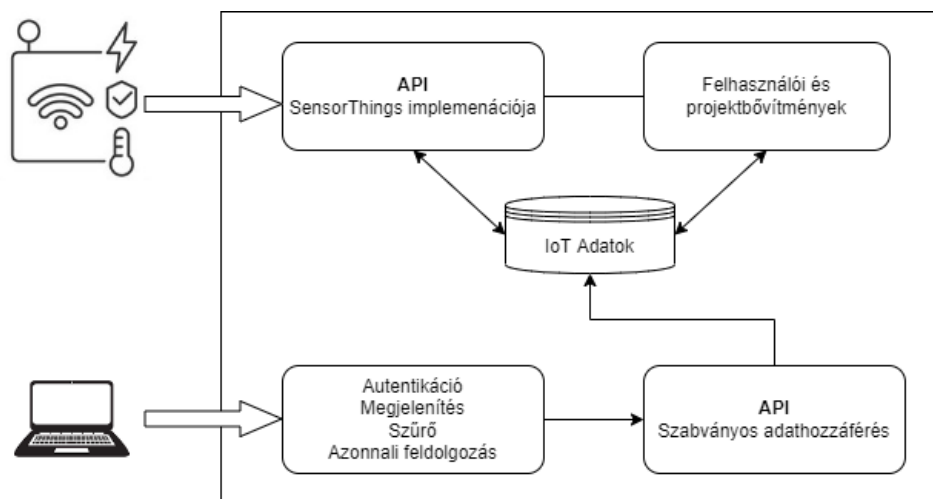
- API adatgyűjtés az internethez csatlakoztatott IoT-eszközökről
- Adat- és adatforrás-kezelési komponens
- Adatfeldolgozás és -elemzési komponens
- Webes felület az IoT-eszközök által gyűjtött adatok elérésére és megtekintésére
- Webes felület adatfeldolgozáshoz és elemzéshez

3.1. Adatgyűjtési és kezelési komponensek

Az adatgyűjtési és kezelési komponensek a következőkért felelősek:

- Adatok lekérése heterogén adatforrásokból, főleg intelligens érzékelőkből, szenzorhálózatokból vagy általános IoT-eszközökből egy webes kéréseket fogadó felületen keresztül;
- Adatforrásokból gyűjtött adatok tárolása;
- Adatkezelés: adatforrások által küldött adatok fogadása, dekódolása és tárolása;
- Adatforrások kezelése: az adatforrások állapotáról információk megőrzése.

Az IoT-eszközök adatokat kapnak a csatlakoztatott érzékelőktől, és továbbítják azokat a cloud-ban telepített adatgyűjtési felületre. Az 1. ábra az IoT-szolgáltatás adatgyűjtési és -kezelési összetevőit mutatja be. Az adatgyűjtést az OGC által definiált SensorThings API [8] szabványos megvalósítása biztosítja, amely lehetővé teszi a HTTP- és MQTT-kéréseket az IoT-eszközökről. Az adatok egy PostgreSQL adatbázisban lesznek tárolva Postgis kiterjesztéssel. A használt adatmodell egy szabványos modell [8], amelyet az OGC határoz meg. Az adatokhoz való hozzáférés szabványos API-n keresztül történik, a felhasználó által kiválasztott konfigurációkon és szűrőkön alapul. Ezen túlmenően az érzékelőktől gyűjtött adatok ezen az API-n keresztül elérhetőek lesznek az ugyanazon szabványnak megfelelő külső alkalmazások számára. A hitelesített felhasználók egy webalkalmazáson keresztül hozzáférhetnek a saját érzékelőikről gyűjtött adatokhoz.

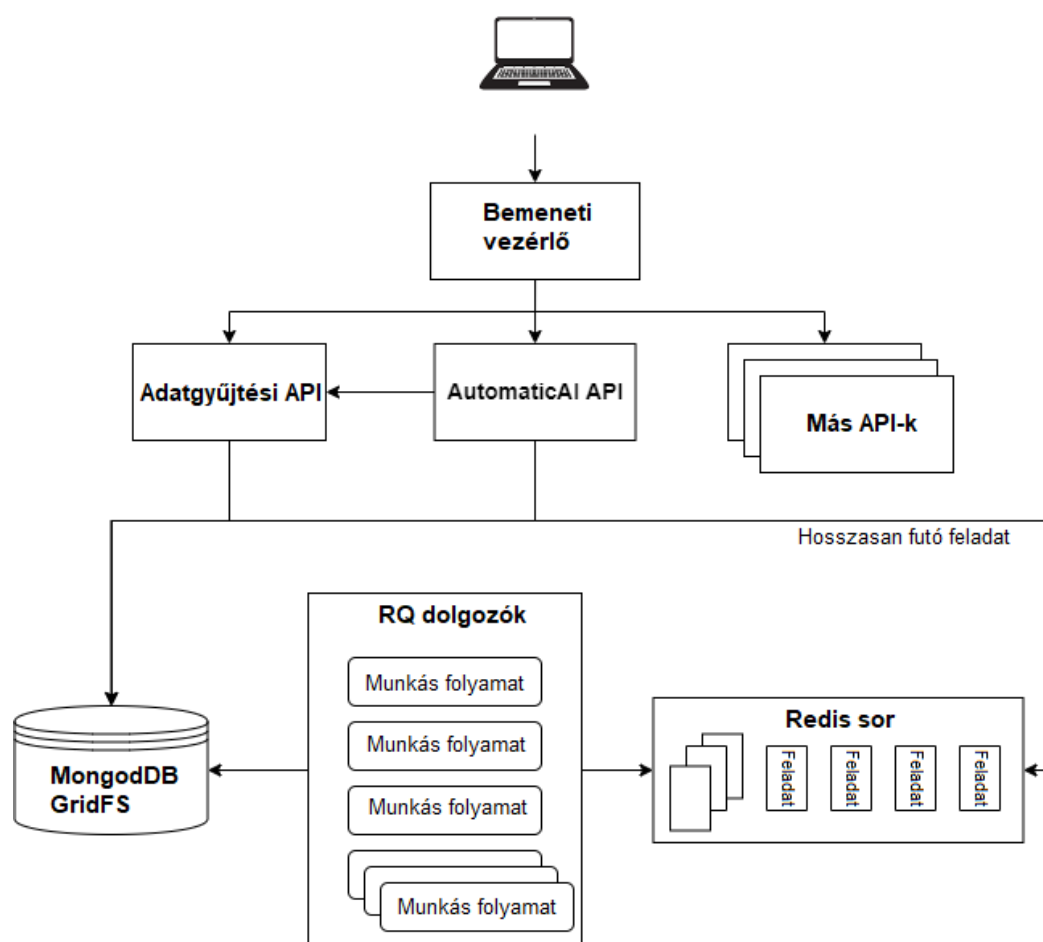


1. ábra. Adatgyűjtési és -kezelési komponensek

3.2. Az adatfeldolgozás és -elemzési komponens

Az adatfeldolgozási és -elemzési komponens az AutomaticAI platformhoz [14] cikkben bemutatott és fejlesztett komponensekre épül. A platform újratervezésére azért volt szükség, hogy a cloud infrastruktúrán keresztül szolgáltatásként elérhetővé váljon. A platform kezdeti komponenseit mikroszolgáltatásokként újracsomagoltuk, és a Dockerrel konténerbe helyeztük. A felhasználói adatok, konfigurációk, felhasználói fájlok és ML modellek egy adatbázisban tárolódnak. A webes felület lehetővé teszi a felhasználók számára az AutomaticAI szolgáltatások elérését, valamint garantálja:

- Adatfájlok feltöltése/letöltése
- Adatok megjelenítése különböző formákban (pl. idősorok, hisztogramok stb.)
- Az adatfeldolgozási és elemzési eljárások lépéseinek sorrendjének meghatározása
- A feldolgozási eredmények megtekintése és mentése
- ML modell beszerzése a kiválasztott adatokhoz, a PSO-SA [12] automatikus kiválasztási mechanizmusának használatával
- Egy korábban megszerzett ML modell újrafelhasználása



2. ábra. Az anomáliák észlelésére szolgáló platform architektúrája

3.3. Anomáliákat osztályozó algoritmusok kiválasztása

Terveztünk egy algoritmust [12], amely automatikusan kiválasztja a legjobb modellt és behangolja hiperparamétereit emberi beavatkozás nélkül, és az anomália-észlelő platform kiegészítő elemeként implementáltuk. Ez az algoritmus a részecskeraj-optimalizálást kombinálja a szimulált hűtéssel, hogy kizárólag a bemeneti adatok alapján keresse meg a legjobb mesterséges intelligencia modellt.

4. KÍSÉRLETEK

A legjobban teljesítő mesterséges intelligencia-algoritmustípus kiválasztására szolgáló automatikus módszerünk [12] és az IoT platformunk tesztelése és értékelése érdekében kutatásunkat a GECCO 2017 [9] adatkészletén végeztük. Ez az adatkészlet több mint 100 000 példát tartalmaz valós adatokra az ivóvíz minőségéről. Az adatkészlet 9 mért attribútumot és egy céloszlopot tartalmaz, mely egy bináris érték, amely azt jelzi, hogy történt-e jelentős változás a vízminőségben vagy sem. A mért attribútumok a következők:

- Tp - a víz hőmérsékletét jelöli Celsius fokban,
- pH - a víz pH-értéke,
- Leit - elektromos vezetőképesség,
- Trueb - a víz zavarossága,
- Redox - Redox potenciál,
- Cl, Cl_2 - klór-dioxid,
- Fm, Fm_2 - Áramlási sebesség.

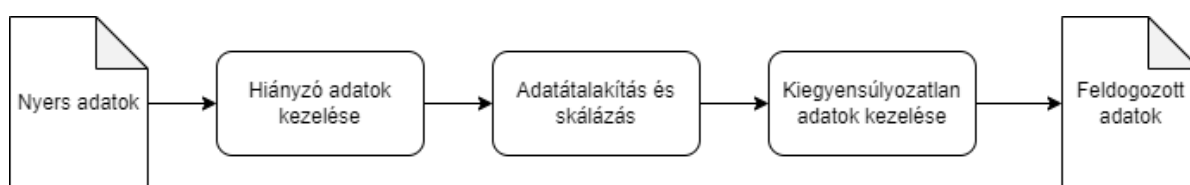
A fent bemutatott értékek nagyon jó vízminőségi mutatók. Például a vízben oldott szilárd anyagok mennyisége határozza meg az elektromos vezetőképességet, tehát ez jó mutatója a víz minőségének. A zavarosság a folyadék relatív átlátszóságának mutatója. A víz zavarossága alapján agyag, iszap, nagyon kicsi szerves és szervetlen anyagok, például mikroorganizmusok mutathatók ki. Ha az ivóvíz pH-járól beszélünk, a pH normál tartománya általában 6,5 és 8,5 között van.

A fő cél ebben az esetben az adatkészlet változásainak/eseményeinek észlelése. Mivel ezek a változások ritkák, az adatkészlet erősen kiegyensúlyozatlan (lásd: 4. ábra, bal diagram), így a probléma anomália-észlelési problémának tekinthető.

4.1. Az adatok előkészítése

Az AI/ML modellek teljesítményének növelése érdekében jó gyakorlat a nyers adatok előfeldolgozása az osztályozási algoritmusok használata előtt. A zaj, a hiányzó adatok stb. negatívan befolyásolhatják az AI-modell teljesítményét.

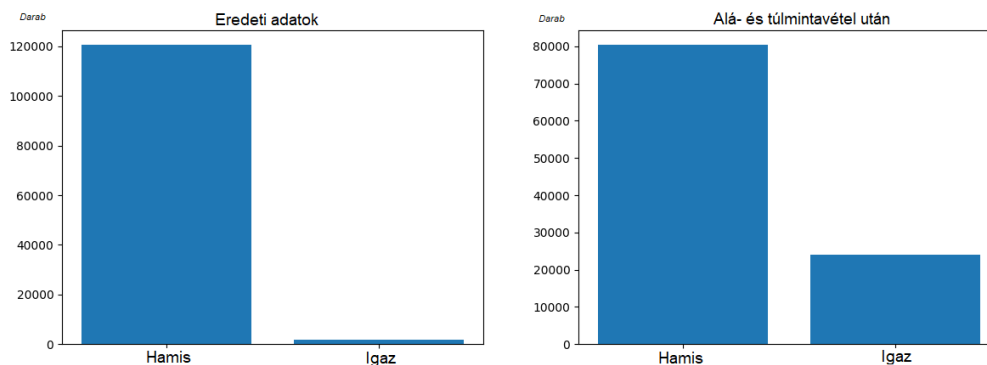
A 3-as ábra azt az adat-előfeldolgozási folyamatot ábrázolja, amelyet az ivóvízminőségi adatkészlet elkészítéséhez használtunk. Mivel az adatkészlet minden oszlopában körülbelül 10 000 hiányzó érték található, feltétlenül szükséges a hiányzó adatok kezelése. Többféle megközelítést is kipróbáltunk, először eltávolítottuk a hiányzó értékeket tartalmazó sorokat, de így túl sok értékes példát veszítettünk el. A következő próbálkozás az volt, hogy az értékeket átlaggal, medián értékkel vagy nullákkal töltöttük ki. A legmegfelelőbb megoldásnak a hiányzó oszlopok nullákkal való kitöltése bizonyult.



3. ábra. Az adatfeldolgozás lépései

Számos algoritmus létezik, például az SVM vagy a Neurális hálózatok, amelyek esetében a különböző oszlopok eltérő skálája negatívan befolyásolhatja a modell teljesítményét. A probléma megoldása érdekében az adatokat különböző transzformációs stratégiákkal átalakítjuk, mint például a min-max skálázás, a [0, 1] intervallumban történő skálázás és a normalizálás. Esetünkben a legjobb eredményeket a [0, 1] intervallumú vagy standard skálázás adta.

Az utolsó lépésben, mivel a bemeneti adatok erősen kiegyensúlyozatlanok, a SMOTE-t (Synthetic Minority Oversampling Technique) [10] használtuk a kisebbségi osztály túlmintavételezésére és a RUID-ot (Random Under-sampling Imbalanced Datasets) [11] használtuk a többségi osztály alulmintavételezésére, így próbálva újra kiegyensúlyozni az adatkészletet. Ezen algoritmusok futtatása után egy közelebbi címkeszámmal rendelkező adatkészletet kaptunk (lásd: 4. ábra).



4. ábra. A túl- és alulmintavétel eredménye (bal diagram - eredeti adatok, jobb diagram – RUID és SMOTE eredménye)

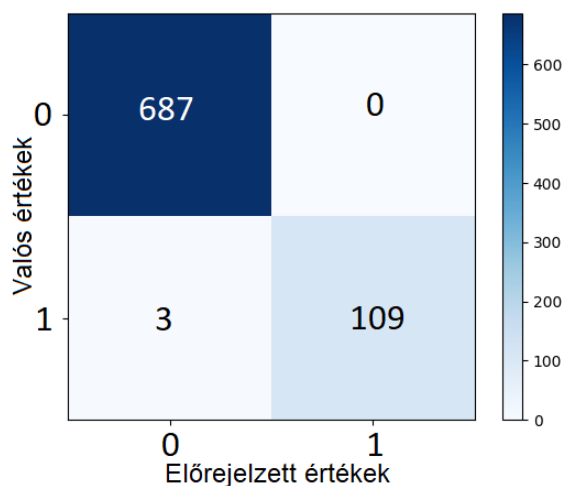
4.2. Kísérleti eredmények

Az előfeldolgozott adatkészletet átadtuk az AutomaticAI platformnak, amely kiválasztotta a legjobb modellt és beállította a hiperparamétereit. A platform különböző algoritmusok összehasonlítására és párhuzamos értékelésére is használható. Több mesterségesintelligencia-algoritmus összehasonlítása után az eredmények az 1-es táblázatban láthatóak.

1. Táblázat. Algoritmusok teljesítménye

ALGORITMUS	F1-MÉRTÉK
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 AGGRESSIVE CLASSIFIER	45.36%
RIDGE CLASSIFIER	37.34%

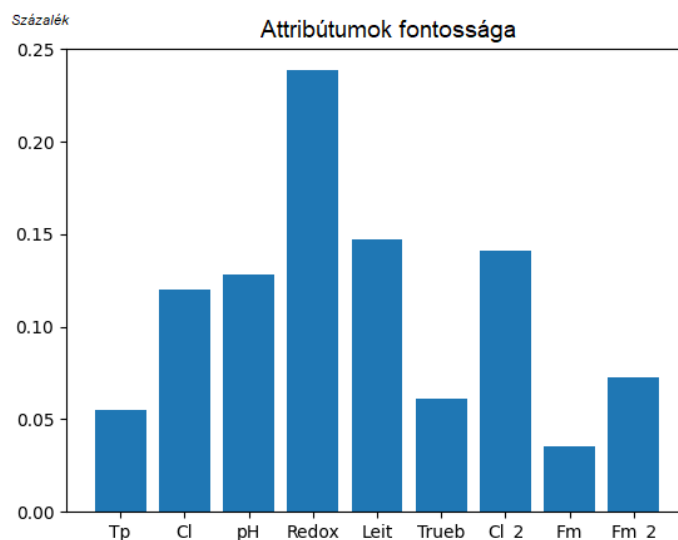
Amint az a 1-es táblázatban látható, a platformnak sikerült automatikusan kiválasztania a modellt és beállítania a hiperparamétereit. A kiválasztott modell a Random Forest volt, 99,93%-os tesztelési F1-mértékkel és 99,49%-os validációs F1-mértékkel (nem látott adatok felhasználásával). A nem látott hitelesítési adatokat használó zavarmátrixa az 5. ábrán látható.



5. ábra. Teszt adatok zavarmátrixa

Számos algoritmus létezik, amelyek hasonló eredményeket értek el, mint például a Random Forest, az Extra Tree osztályozó vagy a döntési fa. Ez normális, mert ezek az algoritmusok nagyon hasonlóak, mindegyik fa alapú ensemble metódus. A Multi Layer Perceptron is magas F1-pontszámot ért el. Ezt a pontszámot úgy kaptuk meg, hogy 9 teljesen összekapcsolt réteget használtunk, 512 csomóponttal kezdve, majd a csomópontok számát 2 hatványával csökkentettük, így a következő rétegekben 256, 128, 64, 32, 16, 8, 4 ill. végül 2 csomópont két osztály osztályozására. Még ha az MLP Classifier magas f1-pontszámot is ért el, az tanulási idő sokkal hosszabb, mint a Random Forest vagy az Extra Tree Classifier esetében, így összességében ennek a versenynek a győztese a Random Forest lett.

A kiválasztott modell eredményeinek magyarázatához és megértéséhez a jellemzők fontosságát használtuk, amely az egyes jellemzők relatív fontosságát jelzi előrejelzéskor (lásd: 6. ábra). Ahogy ezen az ábrán is láthatjuk, az ivóvíz minőségi anomáliák kimutatásának legfontosabb jellemzője vagy attribútuma a Redox, vagy más szóval az Oxidation Reduction Potential, amely azt jelzi, hogy a víz mennyire fertőtlenített vagy szennyezett az oxidációja és redukciója tulajdonságai alapján. Ez egy nagyon jó mérőszám a vízminőség becsléséhez, ezért minden vízminőség-ellenőrző rendszernek ORP-mérőket kell használnia.



6. ábra. Attribútumok fontossága

A jellemzők fontosságának második helyére több attribútumunk van, amelyek nagyon közeli fontossági értékekkel rendelkeznek, például pH, Cl, Cl_2 és Leit. Azért sejtettük ezt az eredményt, mert például a klór-dioxidot (Cl és Cl_2) fertőtlenítőként, valamint szag- és ízsabályozásra használják a vízben és az élelmiszerek higiénijában, így ez jó mutatója lehet a vízminőségnek. A pH-érték is jó mutatója a víz minőségének, mert az ivóvíz pH-értéke 6,5 és 8,5 között kell, hogy legyen. A nagyon alacsony vagy magas pH-érték a víz kémiai szennyezettségét jelezheti. A Leit attribútum a víz elektromos vezetőképessége. A vízben oldott szilárd anyagok mennyisége határozza meg az elektromos vezetőképességet, így ez is jó mutatója a vízminőségnek.

A legkevésbé fontos jellemzők az FM és FM_2 attribútumok. Ez a két attribútum az áramlási sebességet jelzi, ami valójában nem az ivóvíz minőségét jelzi, hanem azt, hogy egy adott időn belül mennyi víz jön ki a csapból. A víz zavarossága szintén alacsony jellemző fontosságú értékkel rendelkezik, ami annak a jele lehet, hogy ez az attribútum redundáns lehet, és kifejezhető az adatkészletben lévő egyéb attribútumok kombinációjával.

5. KÖVETKEZTETÉSEK

Ebben a cikkben röviden ismertettük anomália-észlelő platformunkat, és részletes áttekintést adtunk az architektúráról. Ezt a platformot használtuk az ivóvízminőség anomáliák észlelésének problémájának megoldására. A probléma megoldására először felállítottunk egy előfeldolgozási folyamatot, amelyben a hiányzó adatproblémát több lehetséges megoldás kipróbálásával kezeltük, az adatokat átalakítottuk, majd

SMOTE és RUID segítségével kezeltük a kiegyensúlyozatlansági adatproblémát az osztályozási modell teljesítményének növelése céljából.

A cikk utolsó részében bemutattuk a kísérleti eredményeket, amelyekben több mesterségesintelligencia-algoritmus típus teljesítményét hasonlítottuk össze, és bemutattuk a platformunk által automatikusan kiválasztott és hangolt modell eredményeit.

Az automatikusan kiválasztott algoritmus jól szemlélteti platformunk teljesítményét, hiszen a hibrid PSO-SA algoritmusunk futtatásával minimális emberi beavatkozással 99%-ot meghaladó F1-mértéket értünk el.

Elismerés / Acknowledgment

Ezt a munkát a román Kutatási és Innovációs Minisztérium, a CCCDI – UEFISCDI 452PED/2020 projektszámú pályázata támogatta a PNCDI III keretén belül.

This work was supported by a grant of the Romanian Ministry of Research and Innovation, CCCDI – UEFISCDI, project number 452PED/2020, within PNCDI III.

IRODALMI HIVATKOZÁSOK

- [1] C. Gakii, J. Jepkoech, A classification model for water quality analysis using decision tree, European Journal of Computer Science and Information Technology, Vol.7, No.3, pp.1-8, June 2019
- [2] F. Muharemi, D. Logofătu, and F. Leon, Review on general techniques and packages for data imputation in r on a real-world dataset, Springer, 2018
- [3] A. H. Haghiabi, A. H. Nasrolahi, and A. Parsaie, Water quality prediction using machine learning methods, Water Quality Research Journal, 2018
- [4] Q. T. Boon, Secure Water Treatment SWaT, 2015, <https://itrust.sutd.edu.sg/testbeds/secure-water-treatment-swat/>
- [5] T. C. Hock, Water distribution testbed (WADI), 2016, <https://itrust.sutd.edu.sg/testbeds/water-distribution-wadi/>
- [6] F. Muharemi, D. Logofătu and F. Leon, Machine learning approaches for anomaly detection of water quality on a real-world data set, Journal of information and telecommunication, 2019
- [7] J. Zhang, X. Zhu, Y. Yue, and P. W. Wong, "A real-time anomaly detection algorithm/or water quality data using dual time-moving windows", Seventh international conference on innovative computing technology (INTECH), pp. 36–41, 2017
- [8] OGC SensorThingsAPI, <https://www.ogc.org/standards/sensorthings>
- [9] GECCO 2019 Industrial Challenge: Monitoring of drinking-water quality, <http://gecco-2017.sigevo.org/index.html/Competitions.html>
- [10] C. Nitesh, B. Kevin, H. Lawrence, and W. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research (JAIR), vol. 16, pp. 321-357, 2002, 10.1613/jair.953
- [11] P. Joseph, K. Taghi, D. David, and N. Amri, "Using Random Undersampling to Alleviate Class Imbalance on Tweet Sentiment Data", pp. 197-202, 2015, 10.1109/IRI.2015.39
- [12] Vacariu L., Hangan A., Mocanu M., Pollution Detection On The Cyberwater Platform, in Environmental Engineering And Management Journal, vol. 14, no. 9, pp. 2043-2050, 2015. ISSN: 1582-9596
- [13] Sebestyen G, Hangan A., Czako Z. and Kovacs G., A taxonomy and platform for anomaly detection, 2018 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR), Cluj-Napoca, Romania, 2018, pp. 1-6.
- [14] Zoltan Czako, Gheorghe Sebestyen, Anca Hangan, AutomaticAI – A hybrid approach for automatic artificial intelligence algorithm selection and hyperparameter tuning, Expert Systems with Applications, Volume 182, 2021