Ipari szenzoradatok elemzése statisztikai és regressziós módszerekkel

Analysis of industrial sensor data using statistical and regression methods

FERENCZ Katalin, PhD hallgató[1], dr. DOMOKOS József, egyetemi docens[2], dr. MOLNÁR László, egyetemi adjunktus[2]

[1]Óbudai Egyetem, Alkalmazott Informatika és Alkalmazott Matematika Doktori Iskola, 1034 Budapest, Bécsi út 96/b, tel. +36-1-6665544, ferenczkatalin@yahoo.com
[2]Sapientia EMTE Marosvásárhelyi Kar, Villamosmérnöki tanszék, Koronka, 2 szám, tel. +4 0265 206 210, fax +4 0265 206 211, domi@ms.sapientia.ro, laszlo.molnar@ms.sapientia.ro

Abstract

Industry players must devote significant attention and resources to real-time data processing to timely extract vital information from available datasets. This includes identifying outlier data, filtering fake information, and enabling predictive maintenance through forecasting analysis. This complex analysis process requires the use of various algorithms that support the listed objectives and offer a broad range of solutions. In this study, we describe the application of Apache Spark's integrated system for the statistical analysis of time series data, which significantly accelerates industrial data analysis. Additionally, we present the linear and Random Forest regression models and their achieved results.

Keywords: IIoT, statistical analysis, regression models, time series data, Apache Spark

Kivonat

Az ipar szereplőinek kiemelt figyelmet és erőforrásokat kell szentelniük az adatok valós idejű feldolgozására, hogy képesek legyenek létfontosságú információkat kinyerni a rendelkezésre álló adathalmazokból. Ez magában foglalja a kiugró értékek azonosítását, a hibás információk szűrését, és az előrejelző analízis segítségével a prediktív karbantartás lehetőségét. E komplex elemzési folyamat során szükségessé válik különböző algoritmusok alkalmazása, melyek a felsorolt célokat támogatják, és széleskörű megoldásokat kínálnak. A dolgozatban ismertetjük az Apache Spark rendszer alkalmazását idősoros adatok statisztikai elemzésére, amely felgyorsítja az ipari adatelemzési eljárásokat és bemutatjuk a lineáris és a Random Forest regressziós modelleket és az adatfeldolgozások eredményeit.

Kulcsszavak: IIoT, statisztikai elemzés, regressziós modellek, idősoros adatok, Apache Spark

1. BEVEZETÉS

A mai ipari környezet alapvető hajtóereje a gyors és hatékony adatfeldolgozás és -értékelés. Az ipari szektorban az IoT és IIoT (Industrial Internet of Things) eszközök által generált adatok rohamos növekedése új kihívásokat jelent az adatelemzés terén. Ennek érdekében elengedhetetlen a nagy mennyiségű gyűjtött adatok (szenzoradatok) valós idejű elemzése, hogy időben felismerjük a kiugró értékeket, a hamis adatokat, és előrejelzéseket tehessünk a váratlan költségek elkerülése érdekében. Az adatelemzés által történő anomália detektálás lehetővé teszi a megelőző karbantartás pontos időzítését, ami kulcsfontosságú a hibák megelőzésében és az ipari folyamatok zavartalan működésének biztosításában. Az adatelemzés során számos, a célhoz igazított algoritmust lehetséges alkalmazni, amelyek széles spektrumban mozognak.

Ebben a cikkben bemutatjuk, hogyan lehet az Apache Spark egységes motorjának képességeit felhasználva idősoros adatok statisztikai elemzését végezni, valamint a lineáris és a regressziós

modellek alkalmazását, továbbá a hatékonyságukat is vizsgáljuk az ipari adatelemzési folyamatok felgyorsítása érdekében. Továbbá, az IoT eszközök integrálása által generált hatalmas adatmennyiség kezelése és értékes információvá alakítása jelentős kihívást jelent, amelyre a Random Forest Regression algoritmus alkalmazásával és annak eredményességének vizsgálatával igyekszünk megoldást biztosítani. A cikk célja, hogy átfogó képet adjon az ipari szenzoradatok elemzésére használható statisztikai és regressziós módszerekről, valamint bemutassa ezek gyakorlati alkalmazásait.

2. Valós idejű adatelemzés és regressziós modellek

A regressziós modellek legnagyobb hasznát olyan kontextusokban találhatjuk meg, ahol a vizsgált függő változó folytonos, számos lehetséges értéket felvehet, ilyen példa lehet a hőmérsékletmérés. Ezekben a kontextusokban a regressziós modellek használatának elsődleges célja a magyarázó változó függő változóra gyakorolt hatásának meghatározása [1]. Ezek a modellek képesek egyszerre számos változót kezelni, ezáltal lehetővé téve a változók közötti összefüggések feltárását. Továbbá, a regressziós modellek jelentős hatékonyságot mutatnak az idősoros adatok elemzésében is. Emellett lehetőséget biztosítanak előrejelzések generálására a meglévő adatok alapján [2]. Több regressziós modell is létezik, melyek különböző alkalmazási területeket fednek le:

- Lineáris regresszió: a függő változó és a független változó(k) közötti lineáris kapcsolatok modellezésére használható, minden egyes változó hatását egyértelműen illusztrálja a modell [3];
- Polinomiális regresszió: alkalmazható, ha a független változó és a függő változó között nem lineáris kapcsolat áll fenn;
- Logisztikus regresszió: bináris célfüggő változókhoz használatos, az esélyek logaritmusát modellezi a független változók függvényében;
- Szupport Vektor Regresszió: az alapvető support vector machine elvén alapul és hatékonyan kezeli a nem lineáris kapcsolatokat és a magas dimenziós adatokat;
- Random Forest Regresszió: képes kezelni a nemlineáris kapcsolatokat és a független változók közötti bonyolult interakciókat, amelyhez több döntési fát kombinál [4].

Kutatásunk során részletesebben vizsgáltuk a lineáris és a random forest regressziós modellek implementációját, és értelmeztük eredményeiket a használati esetünk során. Ezenkívül foglalkozunk a regressziós modellek alkalmasával a valós idejű rendszerek esetében.

A statisztikai adatelemzés és az idősor-elemzésre különös hangsúlyt fektetünk a szenzoradatok kezdeti elemzésének kritikus fontossága miatt. Ez magában foglalja a leíró statisztikák alkalmazását, amelyek kulcsfontosságúak az adatok általános tendenciáinak, mint például középérték, medián, szórás, és kvartilisek megértésében. A megvalósítás során foglalkozunk az anomália- és kiugró értékek azonosításával, amely elengedhetetlen a potenciálisan hibás adatpontok vagy rendellenes viselkedés azonosításában. Ezek az elemzési lépések alapvetően fontosak a szenzoradatok viselkedésének mélyreható megértéséhez, amely lehetővé teszi az ipari környezetben dolgozó szakemberek számára, hogy részletes betekintést nyerjenek a gyártási folyamatokba és a berendezések aktuális állapotába.

4. AZ APACHE SPARK ALKALMAZÁSAI AZ ADATELEMZÉSBEN

Manapság az ipari szereplőket leginkább az foglalkoztatja, hogyan tudjanak megfelelni az Ipar 4.0-ra vonatkozóan támasztott elvárásoknak, lépést tartani a technológia gyors fejlődésével, és költséghatékonyan működtetni a mindennapi folyamatokat. Minden ipari egységnek alkalmazkodnia kell az új technológiákhoz és ki kell használniuk a kínált lehetőségeket.

A következőkben bemutatjuk azt a rendszerünket, amely az ipari gyors adatelemzésben és a szenzorértékek idősorainak kiugró értékeinek észlelésében is segítséget nyújt. Egy olyan prototípus rendszeren dolgozunk, amely lehetővé teszi az adatok könnyű tárolását, amelyeket nagyszámú különböző típusú okoseszközből gyűjtünk az Apache Cassandra nyílt forráskódú, elosztott NoSQL adatbázis-kezelő rendszerben [5]. Ezen alapokat használva az Apache Spark, szintén egy nyílt forráskódú adatfeldolgozó rendszer, lehetővé teszi a felhasználó számára, hogy kihasználja a Spark azonnali adatelemzésre való képességét igény szerint. Az Apache Spark által biztosított adatelemzési

lehetőségek magába foglalják a Spark Streaming (majdnem valós idejű) – skálázható és hibatűrő folyamatos adatfeldolgozót; a SparkSQL (Strukturált adatok) – modult a strukturált adatokkal való munka támogatására; az MLlib (Gépi tanulás) – skálázható gépi tanulási könyvtárat és a GraphX (Gráfelemzés) – gráfokra és gráf-párhuzamos számításokra specializálódott modult. Az Apache Spark egy széles körben használt és támogatott nyílt forráskódú eszköz statisztikai elemzésre, gépi tanulásra és különféle adattudományokra. A Spark MLlib könyvtárak használatának célja egy magas szintű és könnyen használható API-készlet biztosítása gépi tanulás és különféle adatelemzési feladatok elvégzésére.

Implementációnkhoz az úgynevezett Combined Cycle Power Plant (CCPP) [6] nyílt forráskódú gyári rendszer adatkészleteit használtuk fel, amely 6 év működését írja le, 9568 adatponttal, és minden adatpont 5 szenzorértéket tartalmaz (környezeti hőmérséklet, környezeti nyomás, relatív páratartalom, vákuum és elektromos energia). Ezt az adathalmazt egy CQL (Cassandra Query Language) parancs segítségével illesztettük be az adatbázisba, amelyet ezután olvashatunk vagy módosíthatunk a Spark-Cassandra közötti összekötő használatával. A Python programozási nyelvet választottuk, így a PySpark interfészt használjuk az adateléréshez.

Az alábbiakban megvizsgáljuk a Spark néhány olyan fontos funkcióját, amelyek lehetővé teszik, hogy könnyen és gyorsan értékes információkat nyerjünk ki a tárolt adatokból, elemezzük adatainkat, észleljük a kiugró értékeket és lineáris regressziós modell segítségével végezzünk előrejelzéseket.

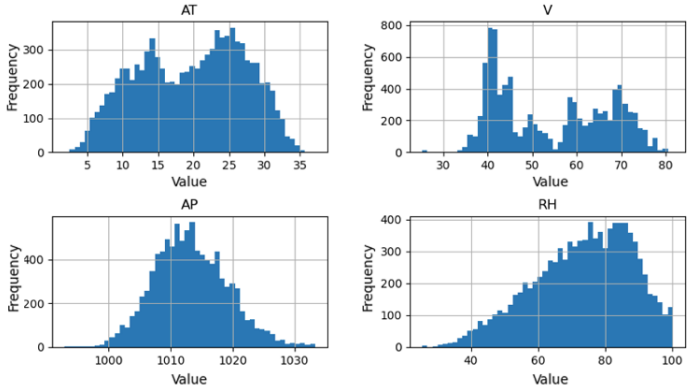
4.1 Az adatok statisztikai elemzése

A Spark beépített funkcióival néhány leíró statisztikát pillanatok alatt kiszámíthatunk. A következő leíró statisztikákat vizsgáltuk meg: átlag, szórás, minimum, maximum, medián, variancia, ferdeség, csúcsosság, kovariancia és korreláció.

1. táblázat: A statisztikai elemzés eredménye

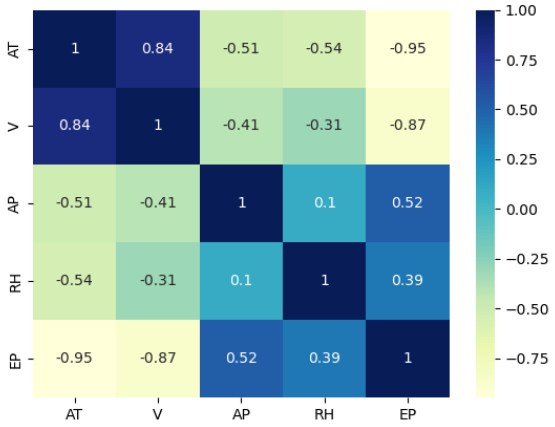
summary	temperature	pressure	humidity	electric power	vacuum
count	9568	9568	9568	9568	9568
mean	19.651	1013.259	73.308	454.365	54.305
stddev	7.452	5.938	14.600	17.066	12.707
min	1.81	992.89	25.56	420.26	25.36
25%	13.51	1009.1	63.32	439.75	41.74
50%	20.34	1012.94	74.96	451.51	52.08
75%	25.72	1017.26	84.83	468.43	66.54
max	37.11	1033.3	100.16	495.76	81.56
median	20.34	1012.94	74.96	451.51	52.08
variance	55.539	35.269	213.167	291.282	161.490
skewness	-0.136	0.265	-0.431	0.306	0.198
kurtosis	-1.037	0.093	-0.444	-1.048	-1.444
covariance – electric power	-120.593	52.546	97.129	-	-188.642
correlation – electric power	-0.948	0.518	0.389	-	-0.869

A statisztikai metrikákon felül a hisztogram ábrázolással lehetséges a frekvenciaeloszlások megjelenítése. A kapott eredmények (1. ábra) alapján a hőmérséklet (AT) és a kipufogógáz vákuum (V) némileg bimodális eloszlást mutat. Továbbá a környezeti nyomás (AP) normális eloszlást követ, míg a relatív páratartalom balra ferdeséget mutat.



1. ábra. Az AT, AP, RH és V hisztogramja

A statisztikai metrikák kiszámítását felhasználva, ha a korrelációs értékeket (2. ábra) a HeatMap segítségével ábrázoljuk, akkor megállapítható, hogy magas korreláció van a hőmérséklet és az elektromos teljesítmény között, valamint magas a korreláció a villamos teljesítmény és a kipufogógáz vákuum között. Továbbá az AP és az RH pozitív korrelációt mutat az EP-vel, míg az AT és a V erős negatív korrelációt mutat.



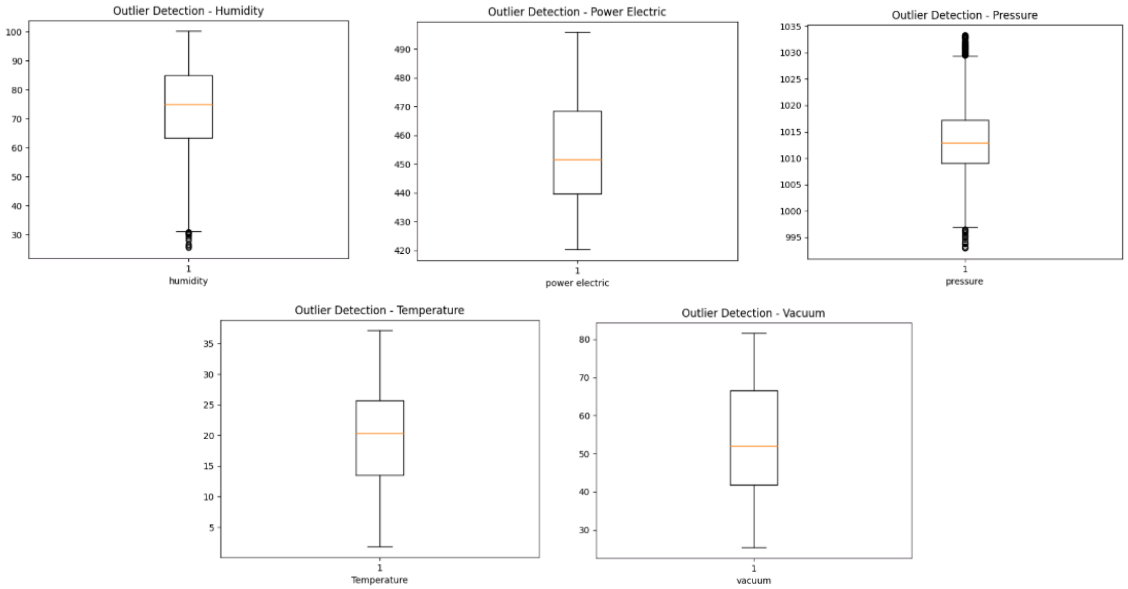
2. ábra. Korreláció HeatMap

4.2 Kiugró értékek detektálása

Az Apache Spark segítségével bizonyos függvények és parancsok futtatásával is észlelhetjük a kiugró értékeket. A teszt során minden érzékelő értékét megvizsgáltuk A feladat az, hogy megtaláljuk az összes rendellenes mérést a DataFrame-ből. Ehhez ki kell számolnunk a felső és alsó küszöbértékeket, amelyek általában 3 szórással esnek távolabb az eloszlás átlagától. A felső határ feletti vagy az alsó határ alatti mérések kiugró értékeknek minősülnek.

$$outlier_{thresholds} = mean \pm 3 * stddev$$

Az eredmények azt mutatják, hogy a kiugró érték észlelése 51 nyomásértéket talált a felső és alsó kiugró küszöbértéken kívül. A többi szenzorérték vizsgálatából azt találtuk, hogy a nyomásértékeken kívül a páratartalom értékek is tartalmaznak kiugró értékeket. A többi szenzorértéknél nem találtunk kiugró értéket. Ha a szenzoradatokat a box plot függvény segítségével ábrázoljuk, akkor a következő diagramokat kapjuk, ahol a fekete pontokkal jelölt kiugró értékeket láthatjuk (3. ábra).



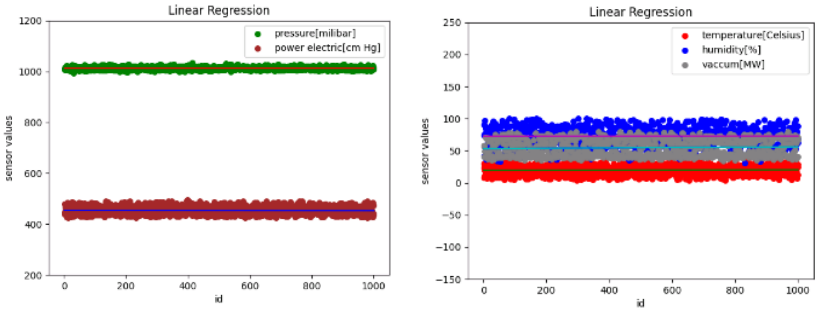
3. ábra. Kiugró érték detektálás – box plot

4.3 Előrejelzés/becslés és regressziós modellek

Az Apache Spark MLlib könyvtárai lehetővé teszik különféle becslések készítését. A lineáris regresszió az egyik legszélesebb körben alkalmazott prediktív modellezési módszer. Az elemzés célja annak ellenőrzése is, hogy a független változó magyarázza-e a függő változót. A következőkben regresszió három megoldását mutatjuk be.

I. Matplot és Numpy

A matplot és a numpy függvénytárak segítségével könnyen megrajzolhatjuk lineáris regressziós egyenesünket és a kiválasztott nyomásértékeket. Annak érdekében, hogy diagramunk értelmezhető és jól áttekinthető legyen, mind az 5 érzékelő értéket egymás mellett ábrázoljuk, így együtt vizsgáljuk őket. Az így kapott grafikonon (4. ábra) jól látható, hogyan helyezkednek el az érzékelők adatpontjai egy egyenes vonalhoz képest.



4. ábra. Lineáris regresszió

II. Lineáris regresszió meghatározása PySpark MLlib segítségével

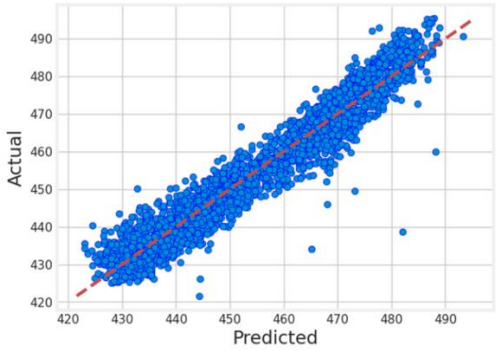
A lineáris regressziós modell felépítéséhez a teljes adatsort, mind az 5 szenzorértéket használjuk. Az elektromos teljesítmény lesz a címke, a másik négy szenzorérték pedig a jellemzők. A jellemzők mind független változók, amelyekről úgy gondoljuk, hogy segítenek megjósolni egy függő változó értékét. A címke egy függő változó, melynek értékét a modellünk megjósolja.

Adatainkat két részre osztottuk: tanítás (70%) és teszt (30%) adatokra. A tanítási adatokat arra használjuk, hogy bizonyos algoritmusok alapján megtanítsuk a modellünket, majd a tesztadatokon végrehajtjuk és ellenőrizzük az előrejelzést. A LinearRegression csomag importálása után prediktív modellezési algoritmusként használjuk a modell létrehozásához. Majd a *fit()* metódus tanítási adatokat felhasználva megvalósítja a modell betanítását. Így a betanított modellünk előrejelzi a tesztadatok értékeit.

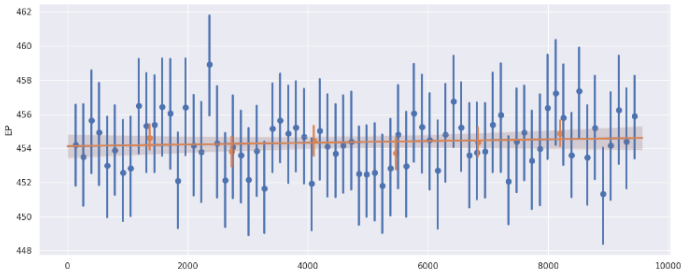
2. táblázat: Az elektromos teljesítmény értékének előrejelzése

prediction	electric power	features
489.94	490.55	[1.81, 1026.92, 76….]
489.88	490.34	[2.34, 1028.47, 69….]
489.56	488.69	[2,58, 1028.68, 69….]
486.27	485.2	[3.0, 1011.0, 80.14…]
482.18	489.38	[3.26, 996.32, 100….]

Egy kiértékelő segítségével ellenőrizzük, hogy modellünk mennyire jósolja meg a címkét, azaz esetünkben az elektromos teljesítmény értékeket, és hogy jó választás volt-e a lineáris regresszió kiválasztása modellünk algoritmusaként. A regressziós modell kiértékelése a Spark ML RegressionEvaluator segítségével történt. A használt mérőszámok az R^2 és az RMSE (négyzetgyök hiba). R^2 értéke 0.926259 és az RMSE értéke 4.75025. Az R^2 értéke nagyban függ attól, hogy hogyan választjuk ki a tanítási és tesztadatokat, így az eredmény ettől függően változhat. Esetünkben az R^2 megközelíti az 1-et, tehát a modell jól illeszkedik és a becsült adatok közel állnak az elvárt értékhez. Ez azt jelenti, hogy a jellemzőként kiválasztott érzékelőértékek magyarázzák és befolyásolják a címke (elektromos teljesítmény) értékeket. Az RMSE a modell által becsült értékek és a tényleges értékek közötti különbségeket méri.



5. ábra. Az előrejelzett és a teszt adatkészletek ábrázolása

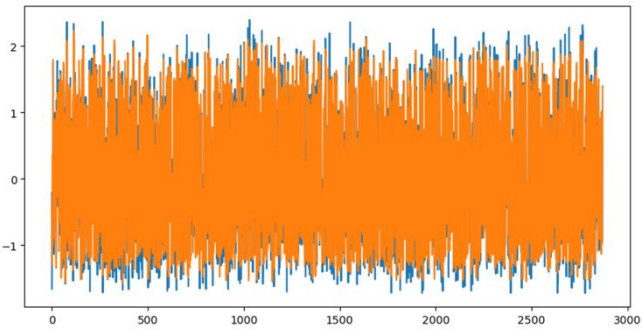


6. ábra. Regressziós diagram

Ha a lineáris regressziós modell során kapott előrejelzett értékeket a teszt adatsorhoz viszonyítva ábrázoljuk, amelyet a regplot függvény segítségével választottunk le az eredeti adathalmaztól, akkor az 5. ábrán látható eredményt kapjuk. Megfigyelhető, hogy a villamos teljesítmény előrejelzett értékei többnyire a modell által felállított egyenes közelében találhatók. Továbbá a regressziós diagram (6. ábra) megrajzolásával megvizsgálható, hogy a havi vagy éves elemzéshez milyen tendenciák figyelhetők meg, hiszen az adatsor és a hozzá tartozó dokumentáció nem tartalmaz információt az időbélyegekről, így az adatok elhelyezkedésea 6 éves időtengelyen belül nem határozhatók meg pontosan.

III. Random Forest Regresszió

A lineáris regresszióhoz hasonlóan az adathalmazra a Random Forest Regression technikát is alkalmaztuk, így lehetőség nyílik regressziós és osztályozási feladatok elvégzésére is, vagyis több döntési fa kombinálásával előrejelzéseket készítenek. A modell jelentősége abból fakad, hogy képes mérni és rangsorolni a bemeneti jellemzők fontosságát a célváltozó előrejelzésében. A teszt és a becsült értékek együttes ábrázolásával láthatjuk, hogy a két értékkészlet nagyon magas lefedettséget mutat.



6. ábra. Az előrejelzett értékek és a tesztadatkészlet ábrázolása a Random Forest regresszió eredményeként

A kapott eredményekből látható, hogy a Random Forest Regression modellünk jobb teljesítményt mutat, mint a lineáris regresszió. A 0,956-os R^2 érték azt jelenti, hogy az EP célváltozó eltérésének 95,6%-a magyarázható a modellel, és az RMSE (0.209) is kisebb értéket ad. Ezért megállapítható, hogy a Random Forest Regression alkalmazása hatékonyabb lehet a lineáris regresszióhoz képest.

5. KÖVETKEZTETÉSEK

Az ipari szenzoradatok elemzésének kutatása rámutat a statisztikai és regressziós módszerek, valamint az adatfeldolgozás és elemzés fontosságára az ipari adatfeldolgozás terén. Ezek a módszerek integrálása lehetővé teszik a vállalatok számára, hogy mélyreható betekintést nyerjenek a gyártási folyamatokba, előrejelezzék a lehetséges hibákat, és optimalizálják a karbantartási ütemezést. A kutatások kiemelik az adatelemzési módszerek folyamatos fejlesztésének szükségességét, a leíró statisztikai paraméterek kinyerését, a kiugró értékek észlelését és a valós idejű ipari rendszerekbe való

integrálás fontosságát, továbbá az alkalmazott algoritmusok eredményeinek szemléltetését az Ipar 4.0 kihívásainak kezelése érdekében. Ezért az Apache Spark nyílt forráskódú adatfeldolgozó rendszer integrálása valós idejű ipari rendszerekbe kulcsfontosságú lépés lehet, amely tovább erősíti a gyártási folyamatok elemzésének és a karbantartási stratégiák fejlesztésének képességét, lehetővé téve a vállalatok számára, hogy hatékonyabban reagáljanak az ipari kihívásokra és előrejelzésekre.

A továbbfejlesztési lehetőségekként fókuszálunk az Apache Spark GraphX és Streaming moduljainak integrálására, amelyek új lehetőségeket nyitnak meg a valós idejű adatfeldolgozás és a komplex hálózati adatstruktúrák elemzése terén. Egy intuitív felhasználói felület kifejlesztése segíthet a rendszer könnyebb kezelhetőségében és az eredmények gyorsabb interpretálásában. Továbbá, a kutatás kiterjesztése többféle idősoros adathalmaz integrálására és új adatfeldolgozó algoritmusok bevezetésére növelheti a prediktív modellek pontosságát és alkalmazhatóságát különböző ipari környezetekben.

KÖSZÖNETNYILVÁNÍTÁS

A jelen munkát Magyarország Collegium Talentum programja támogatta.

HIVATKOZÁSOK

[1] Forkuor, Gerald, et al. "High resolution mapping of soil properties using remote sensing variables in south-western Burkina Faso: a comparison of machine learning and multiple linear regression models." PloS one 12.1 (2017): e0170478.

[2] Ali, Iftikhar, et al. "Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data." Remote Sensing 7.12 (2015): 16398-16421.

[3] Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[4] Coulston, John W., et al. "Approximating prediction uncertainty for random forest regression models." Photogrammetric Engineering & Remote Sensing 82.3 (2016): 189-197.

[5] Ferencz, Katalin, and József Domokos. "Rapid Prototyping of IoT Applications for the Industry." 2020 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR). IEEE, 2020.

[6] https://archive.ics.uci.edu/ml/datasets/combined+cycle+power+plant