



TDK-DOLGOZAT

Bodor Tamás Levente

Bodor Tamás Levente, mérnökinformatikus MSc 2023-2025. (3. féléves)

INTEGRÁLT IRÁNYÍTÓRENDSZER JELEINEK **PROBLÉMAVEZÉRELT VIZSGÁLATA GÉPI TANULÁSSAL**

Témavezető: Dr. Dulai Tibor, egyetemi adjunktus

Pannon Egyetem,

Műszaki Informatikai Kar,

Rendszer- és Számítástudományi Tanszék

Pannon Egyetem

2024. december 25.



1 Tartalom

1	Tartalom	3
2	Bevezető.....	5
2.1	Előszó	5
2.2	Fontos kifejezések	6
2.3	Kontextus.....	7
2.4	Az adat.....	8
2.5	Kihívások.....	8
3	Tartalmi tárgyalás.....	10
3.1	Vizsgálat tárgya	10
3.2	Módszerek	11
3.2.1	Adat vizsgálata	11
3.2.2	Adat előkészítése.....	15
3.2.3	További elemzés.....	19
3.2.4	Modell építés	20
3.2.5	Modell kiértékelés	21
4	Összegzés	26
4.1	Értékelés	26
4.2	További megfontolások	26
4.3	Köszönetnyilvánítás.....	28
4.4	Hivatkozások	29
5	Adatok és vizualizációk.....	30
5.1	Páros eloszlások.....	30
5.1.1	Kiegyenlítés nélkül.....	30
5.1.2	SMOTEENN kiegyenlítéssel	32
5.2	Idősor összehasonlítások és konfúziós mátrixok.....	33
5.2.1	Gaussian Naive Bayes osztályozó modell.....	33
5.2.2	LinearSVC osztályozó modell	35

5.2.3	AdaBoost osztályozó	36
5.2.4	RandomForest osztályozó	37
5.2.5	HistGradientBoosting osztályozó	38
5.2.6	KNeighbors osztályozó	38
5.2.7	LogisticRegression osztályozó	39
5.3	Táblázatok	40
5.3.1	A jobbnak ítélt tanított modellek paraméterei és metrikái	40

2 Bevezető

2.1 Előszó

Ebben a kutatásban egy korábbi munkámat folytatom, illetve egészítem ki. Abban a kutatásban a szennyvíztisztító telep anaerob rothasztó alrendszerének zaját lényegi feldolgozás nélkül vettem alapul egy offline megerősítéssel tanuló ágens tanítására, mérsékelt sikeres eredménnyel. A vizsgálat tárgya az volt, hogy további ismeretek bevitele nélkül, pusztán a jutalmazó-büntető rendszer, illetve a historikus adathoz való hasonlóságra szorítás elég-e arra, hogy az ágens egy operátornak megfelelő szinten elsajátítsa a *habzás* nevű veszélyes, és üzemmenetet akadályozó állapot megszüntetését. A mérsékelt sikerből adódott, hogy a problémát más megközelítésekkel is megvizsgáljam, illetve ha már így teszek, akkor egy olyan értékelési módszert alkalmazzak, amivel akár a megerősítéssel tanuló ágens viselkedésével is összehasonlíthatom a rendszert.

Jelen munkámban ennek a másik megközelítésnek a kidolgozását végzem el azokkal a kikötésekkel, hogy az adatsor, amelyet felhasználok a célra, ugyanarról a rendszerről származik (bár időhorizontját tekintve jócskán kibővített) és kísérletet teszek egy univerzálisan alkalmazható összehasonlítási módszer megalkotására.

A munkám gyakorlati jellegű célja, hogy egy figyelmeztetéseket adó szoftverasszisztens jöjjön létre, tulajdonképpen egy soft-szenzor, amely az említett veszélyes állapot felismerésében segíti az operátort a mindennapokban.

2.2 Fontos kifejezések

trendváltozó: A technológiai rendszer azon változói, amik percenkénti gyakorisággal megvizsgálásra kerülnek, és ha az értéke megváltozott a legutóbbi érték óta, akkor az új érték archiválásra kerül. Az összes itt feldolgozott adat ilyen trendváltozóból származik, aminek az oka, hogy ahhoz nyertem hozzáférést.

éjféli törlés: Az egyes üzemidők tárolása az üzemállapotnál stabilabb módon ad képet az egyes üzemállapotokról a rendszerben (rövid jelkimaradás jellegű hibák az értéktartás miatt nem zavarják a változó általános helyességét), illetve a beavatkozások lefutása könnyen összehasonlítható az előiránnyal. Az üzemidők azonban az általános üzemű berendezések esetén túlsordulnának, ezért időnként nullázni kell a változó értékét. Ez gyakorlatias okokból minden nap éjfélikor történik, így minden üzemidő legfeljebb 24 órányi értéket tárolhat.

műtárgy: ebben a kontextusban az anaerob rothasztás helyszínéül szolgáló gázfejlesztő kamra vagy reaktor

CSV: *Comma-separated values*, vagyis vesszővel elválasztott értékek. A táblázatos jellegű adatok egyik lehetséges kódolási módszere. A gyakorlatban az értékek a fájlban nem szükségszerűen vesszővel vannak tagolva.

soft-szenzor: olyan szoftver, ami több mérési értékből számítások segítségével állít elő olyan jelet, amit így nem szükséges külön mérőeszközzel vizsgálni.

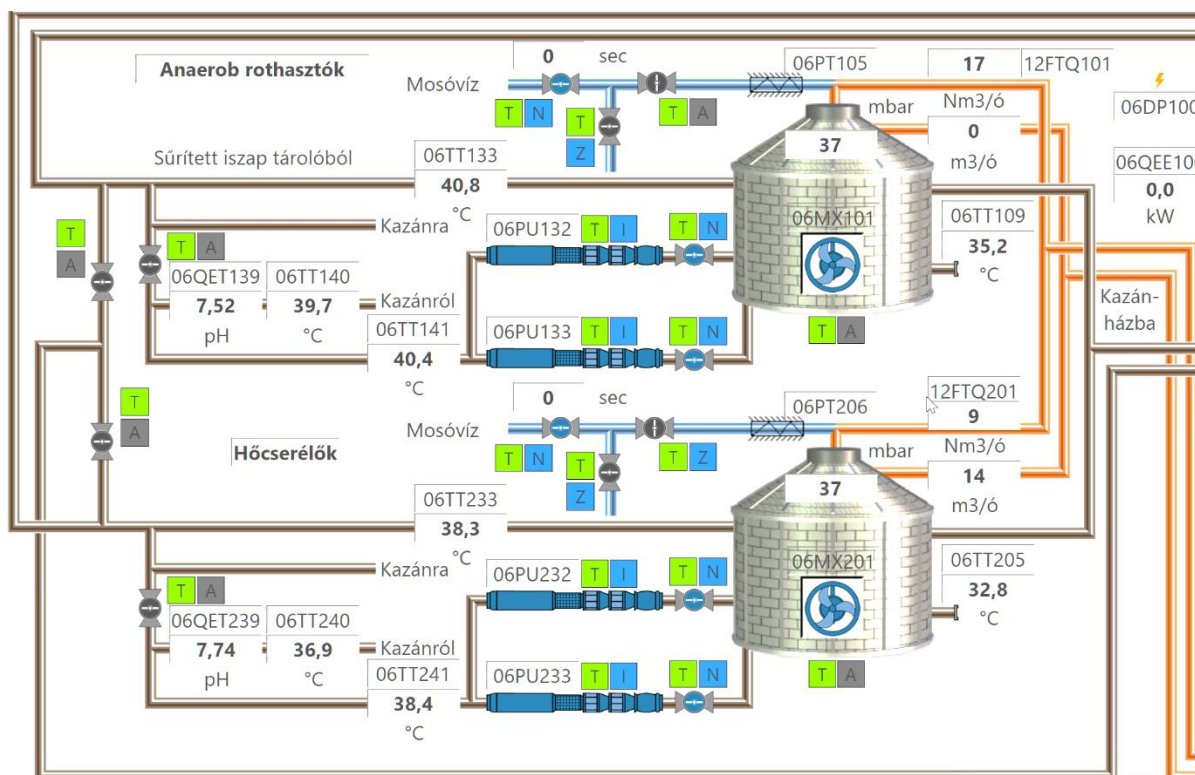
feature selection: a gépi tanulás számára hasznos, jelentéssel bíró bemeneti jelek és változók kiválasztásának folyamata.

klasszifikáció vagy **osztályozás:** olyan eljárás, ami során az adatokat előre meghatározott csoportokhoz rendelünk hozzá hasonlóság alapján.

zaj: valamilyen zavaró körülmény vagy az annak hatására a jelen megjelenő érték vagy értékváltozás

2.3 Kontextus

A szennyvízfeldolgozás során keletkező szennyvíziszap anaerob bomlása során biogáz fejlődik. A szennyvíziszapban jelenlévő organizmusok azonban az iszap habzását okozzák. Ilyenkor a zárt műtárgyban, illetve a recirkulációs rendszerben a szennyvíziszap nagy térfogatú habot nyer, ami a gázzal együtt a gázvezetékbe juthat, a gázrendszerben robbanásveszélyes állapot előidézve.



Az anaerob rothasztó (biogáz termelő) alrendszer technológiai ábrája

A veszélyes állapot megelőzésére a folyamatirányító az irányított rothasztást leállítja, a habzást lemosó vízbeeresztéssel megszünteti, a műtárgy iszaptartalmát csökkenti.

A folyamatirányító rendszerre jellemző, hogy a beavatkozást megelőzően, a habzás során a szenzorok hibás értékeket adnak, mivel nem alkalmasak a habzó iszap fizikai tulajdonságainak mérésére. Ezek a hibás adatok jellemzően kis mértékű, még validációs korlátokon belüli értéket adnak, viszont a hibázás véletlenszerűsége miatt „ugráló” jelet eredményeznek. Ez a zaj az operátor számára jól követhető, de az irányító rendszer nem végez statisztikai szűréseket a jel tulajdonságainak megismerésére.

Az operátor gyakorlati tapasztalata szerint az ugrálás mértéke egyértelműen megnő a habzással terhelt időszakban. Ebből az összefüggésből kiindulva kézenfekvő, hogy a zaj mértékét, mint különálló mérési értéket vizsgáljuk, és ebből következtessünk a habzás jelenlétére.

2.4 Az adat

A kutatásom gyakorlati alkalmazhatóságára tekintettel kifejezetten a gyakorlati világból származó adatok felhasználását tűztem ki célul, ezért a vizsgálat során ragaszkodtam ahhoz, hogy egy szennyvíztelep irányítórendszeréről szerzett adatokon végzem a kellő átalakításokat.

Az adat forrása a zalaegerszegi szennyvízkezelő telep anaerob rothasztó alrendszerének irányítórendszer-archívuma. A kutatásomat megalapozta, illetve megelőzte a (részben jelen dolgozatomban is tárgyalt) adatmegismerés, illetve *feature selection*. Jelen kutatásomban már csak a diszkrét differenciálásra szánt, illetve a beavatkozási állapotot jelző változókat használtam fel, a korábbi vizsgálatok eredményét figyelembe véve.

Az itt tárgyalt vizsgálatokhoz 2019. január 1. éjféltől a mentés időpontjáig, azaz 2024. szeptember 30. reggel 07:55-ig percenkénti felbontással regisztrált adatot használtam fel. Az archívum az első teljes adatpontot (ahol már minden jel rendelkezik értékkel) 2020. április 6-án 11:03 percnél tartalmazza. A vizsgálatához a korábbi, hiányos adatpontokat eldobtam.

Az adatot évenkénti bontásban, egy nem szabványos CSV fájlba exportálva kaptam meg, ezen kezdetben szöveges átalakításokat végeztem, hogy a kutatás során használt *pandas* csomaggal kezelhető legyen.

Az egyes változók értékei csak egy (számomra ismeretlen, és akár változónként különböző) küszöbszint feletti változaskor frissülnek, ezért a közbeeső hiányokat előreterjesztés módszerrel pótoltam. (*pandas.DataFrame.fill*)

2.5 Kihívások

A kutatás alapját az az elgondolás adja, hogy a mérések zajjal való terheltsége szoros összefüggésben áll a műtárgyban tapasztalható habzással. Az első kezelendő kihívás ebből a feltételezésből adódik. Bár az operátor empirikusan igazolta – zajjal terhelt jelhez emlékezetből hozzárendelte a habzási eseményt –, hogy a feltételezés megalapozott, nem oszlatta el annak a lehetőségét, hogy a zajjal terhelt jel más okból is bekövetkezhet, illetve nem tudta konkrét mértékhez kötni a tapasztalást. Tehát állíthatjuk, hogy minden habzás a jel bizonyos (ismeretlen) mértékű zajosodását eredményezi, de nem állíthatjuk, hogy a jel zajossága, vagy annak mértéke egyértelműen habzás jelenlétét implikálja.

A historikus adatok megismerése során az is egyértelművé vált, hogy a jelek nem pusztán zajosak, de hiányosak is. Ez két jellemző okból adódik. Az egyik tervezési eredetű: a trendváltozók archiválása csak a változóérték megváltozásakor történik. A megváltozást százalékos arányban kell érteni, így akkor, amikor a jelérték csak minimálisan változik meg, nem történik mentés. Ez a gyakorlatban akár több időegységnyi (perc) kimaradást is eredményezhet, amikor a változás mértéke a megadott küszöbérték fölé kerül. Ez a vizsgálatunk szempontjából nem okoz különösebb problémát, mivel kifejezetten a nagyobb megváltozásokat keressük a szoftverszenzor megalkotásához. A második ok az adatátvitel megghiúsulásából fakad, ami lehet adó- vagy vevő oldali kommunikációs hiba is. Ez a jelen akár több órányi, vagy több napnyi kimaradást is jelenthet, attól függően, hogy a hibát mikor sikerült azonosítani

és elhárítani. Ez a típusú hiba már érdemben befolyásolhatja a modell helyességét, azonban fontos hozzátenni, hogy adó oldali hiba esetén az operátor egyébként sem értesül a jelérték szélsőséges megváltozásáról, így amennyiben ez a hiba nem extrém módon áll elő (extrém alatt értendő, ha sok habzási esemény és beavatkozás történik a szenzorok kommunikációs hibája közben, vagy sok habzás hatására történik olyan beavatkozás, amely beavatkozás nem kerül rögzítésre a rendszerbe), úgy feltételezhetjük, hogy olyan nem tökéletes modellt taníthatunk a rendelkezésre álló adatokon, ami képes normális üzemi körülmények között az elvárható szinten jelezni egy valószínű habzást.

A konkrét vizsgálati környezetben kihívást jelent az online mérők szükségessége is. Kutatásom célja kifejezetten az ilyen, kevés mérővel ellátott anaerob rothasztó vizsgálata, de fontos leszögezni, hogy az anaerob rothasztók adatvezérelt irányításának tárgykörében a szakirodalom több online kémiai mérő jelenlétét feltételezi, így a kémiai oxigénigény, az illékony zsírsavkoncentráció, a szervetlen szén, erős ionok, illetve a fejlődő gázban a szén-dioxid koncentrációja (Kazemi et al., 2020), valamint a szennyvíziszap kémhatása (Steyer et al., 1997).

A munkám jelentős részben a többváltozós idősorok adatpontjainak osztályozásos szétválasztásának alkalmazásán alapul. Ehhez azonban nem álltak rendelkezésre címkézett adatok, mivel az irányítórendszer nem regisztrálja a habzást, mint kitüntetett állapotot. Ezt a hiányosságot egy megközelítő módszerrel hidaltam át, amit a módszer tárgyalásában részletesen is kifejték. A módszer alapját az a nem teljesen helytálló feltételezés adta, hogy a habzásra való beavatkozás a habzás jelenlétét feltételezi az azt megelőző időszakra. Mint látható lesz, ez a feltételezés több szempontból is elnagyolt: a beavatkozásról nem állíthatjuk, hogy kizárólag habzásra adott válaszként értelmezhető, nem állíthatjuk, hogy minden habzási eseményre megtörtént a szükséges válasz, illetve nem állapítható meg egyértelműen, hogy milyen hosszúságú időszakot jelent a „megelőző időszak”: jelenthet percek, órákat, vagy akár egy-két napot is, az arra adott válasz időszerezése függ a nem látható üzemmeneti körülményektől, mint hétvége, karbantartás, operátorváltás, illetve erősen változhat az éppen aktív operátor személyétől is.

3 Tartalmi tárgyalás

3.1 *Vizsgálat tárgya*

A kutatás célja, hogy az ismertetett ipari folyamat operátora számára hasznos információkkal szolgáljon a jelekből kinyerhető információk ismertetésével, illetve választ adjon arra a kérdésre, hogy a szűkös mérési értékekből lehetséges-e megalkotni olyan soft-szenzort, ami figyelmeztetést ad az operátornak abban az esetben, ha a jeleken mérhető zaj a korábban habzáskor tapasztalttal hasonló jelleget mutat.

A kutatás célja továbbá, hogy megalkosson egy olyan általánosítható összehasonlítási módszert, amellyel a probléma különböző megközelítésű megoldásai, mint az osztályozó algoritmus illetve a megerősítéses tanuló algoritmus alkalmazásának eredménye egymáshoz viszonyítható és értékelhető.

A kutatás kimondott célja, hogy megalapozza a soft-szenzor tervezését a tapasztalati ismeretek és az archív, nagy felbontású adat vizsgálatával.

3.2 Módszerek

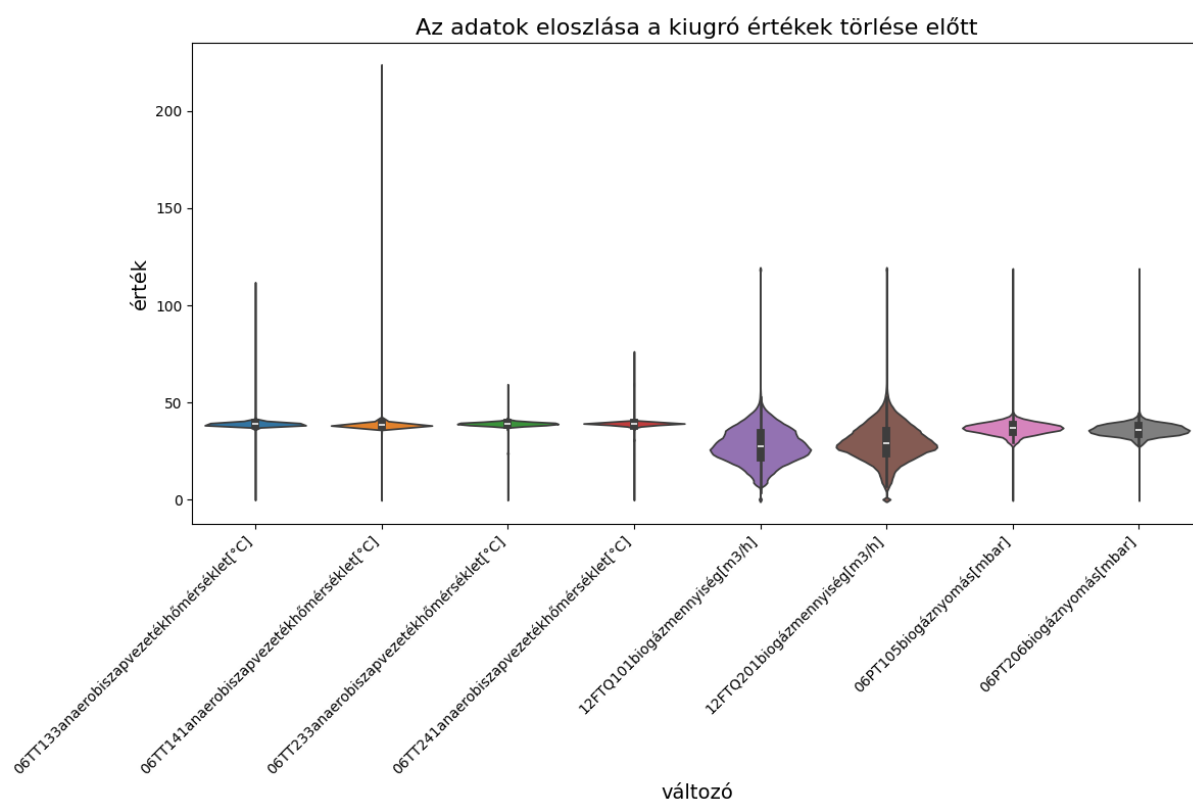
A munkámat a CRISP-ML keretfolyamat (Studer, 2021) lépései szerint szerveztem, ezért ebben a szervezésben mutatom be módszerem lépéseit is.

3.2.1 Adat vizsgálata

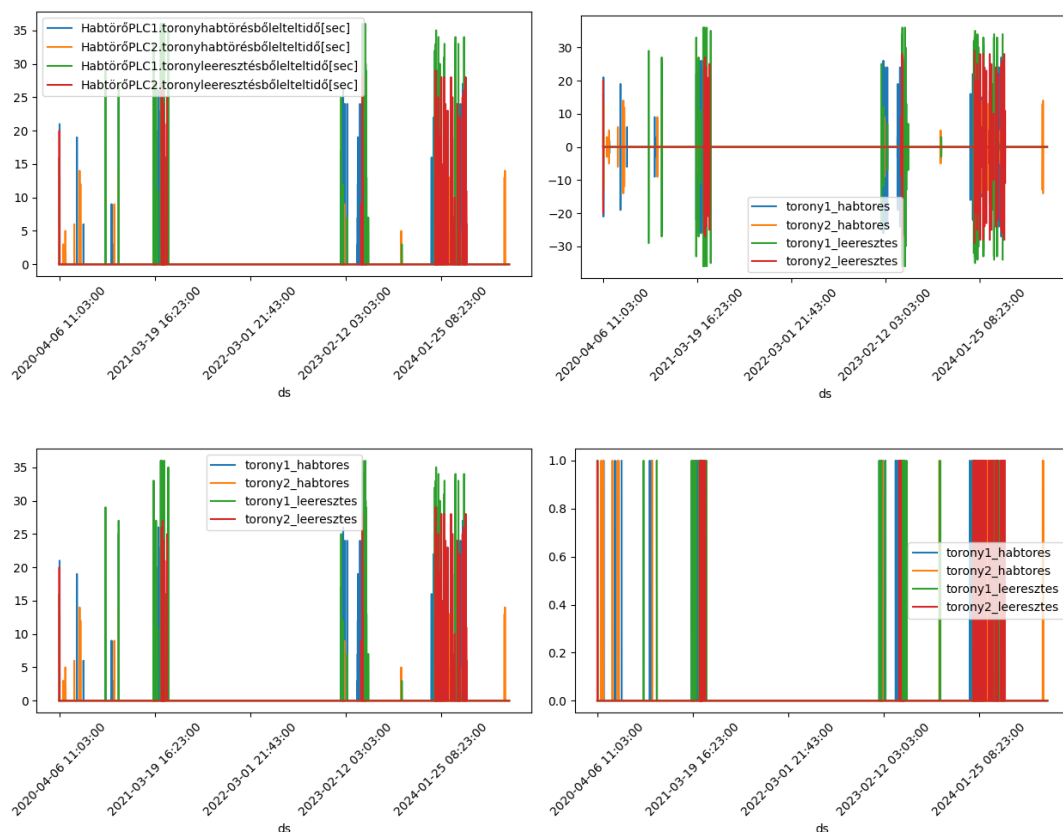
Első lépésként szükséges volt az adatot újra formázni (a nem szabványos CSV formátum miatt), újra mintavételezni (ismétlődő és kimaradó időbélyegek miatt), a darabokat (egyes éveket) összeilleszteni, a hiányzó adatpontokat kitölteni (előreterjesztéses módszerrel).

A vizsgálat körét leszűkítettem a technológiai alrendszer jeleire, amelyek közvetlen összeköttetésben vannak a rothasztott szennyvíziszap és a fejlődő (metán tartalmú) gáz fizikai tulajdonságaival.

Az archívumban tárolt jelek közvetlenül a szenzorokról származó nyers, tisztítatlan jelek, ezért jelentős kiugrások tapasztalhatók:



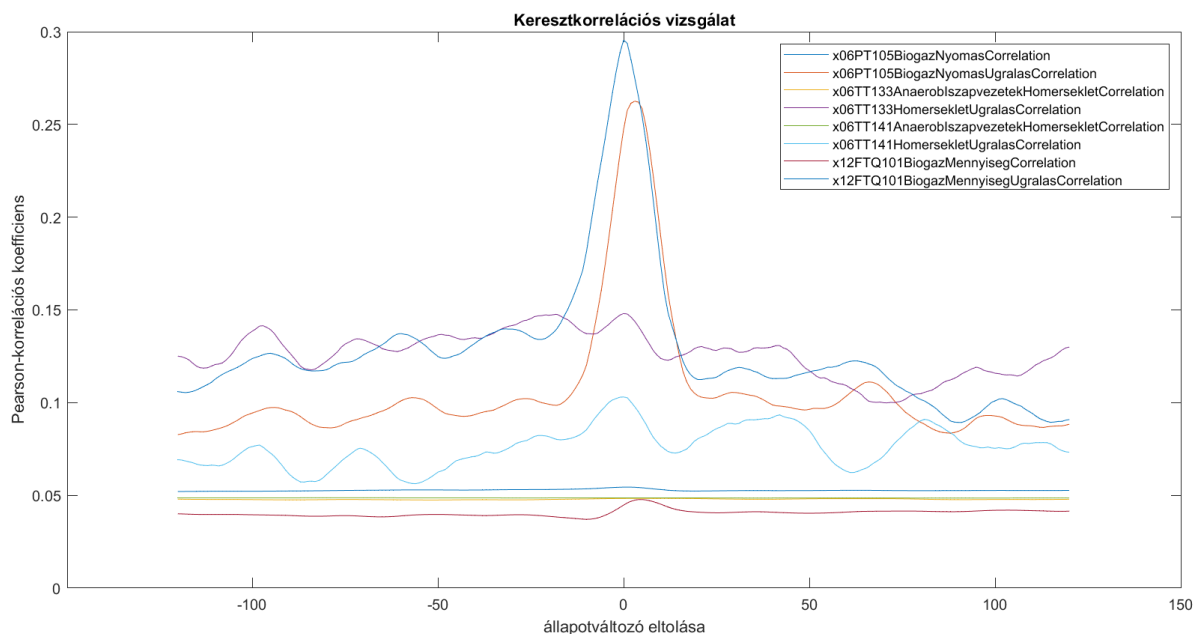
Az üzemi változók (habtörés és leeresztés) integrál jellegűek, a beavatkozások alatt eltelt időt összegzi. A jeleket a beavatkozás megszűnésén túl az éjféli törlések is azonnal 0 értékre változtatják. A jelek differenciálásával, és a negatív differenciák törlésével (az érvényes jelek visszaterjesztésével) egy üzemjelet hoztam létre, ahol a beavatkozást a pozitív értékek, a beavatkozás nélküli időszakot a 0 érték jelzi. Ezekből a jelből azután logikai értéket hoztam létre. Ezzel az eljárással minden pillanatban egy logikai változó segítségével eldönthető, hogy történik-e éppen valamilyen beavatkozás a rendszeren.



Az beavatkozási üzemi állapot jelek megalkotásának lépései

A szenzorokról származó jeleket megtartottam, illetve a diszkrét differenciájukat vettem $h=1$ értékkel, azaz a lehető legnagyobb felbontással alkottam meg a jel megváltozásából és a mérési értéken megjelenő zaj összesített hatását a jelre. A jelnek ezután vettem az abszolútértékét minden időpillanatban. Fontos megjegyezni, hogy a folyamatoknak rendes üzemmenet során stacionárius körülmények között kell történniük, ezért az így megalkotott jelekben nagyságrenddel nagyobb súllyal szerepel a zaj, mint a körülmények valódi megváltozása.

Az így megalkotott jeleket keresztkorrelációs vizsgálatnak vettem alá $h:[-120;+120]$ intervallumon. A vizsgálat első operandusa a vizsgált jel, a második minden esetben a habtörési beavatkozási jel volt. A jobb vizsgálhatóság érdekében a vizsgálatához a pillanatszerű üzemjeleket egy csúszóablakos átlagolással simítottam. Az így kapott Pearson korrelációs koefficiensek láthatók az alábbi ábrán:



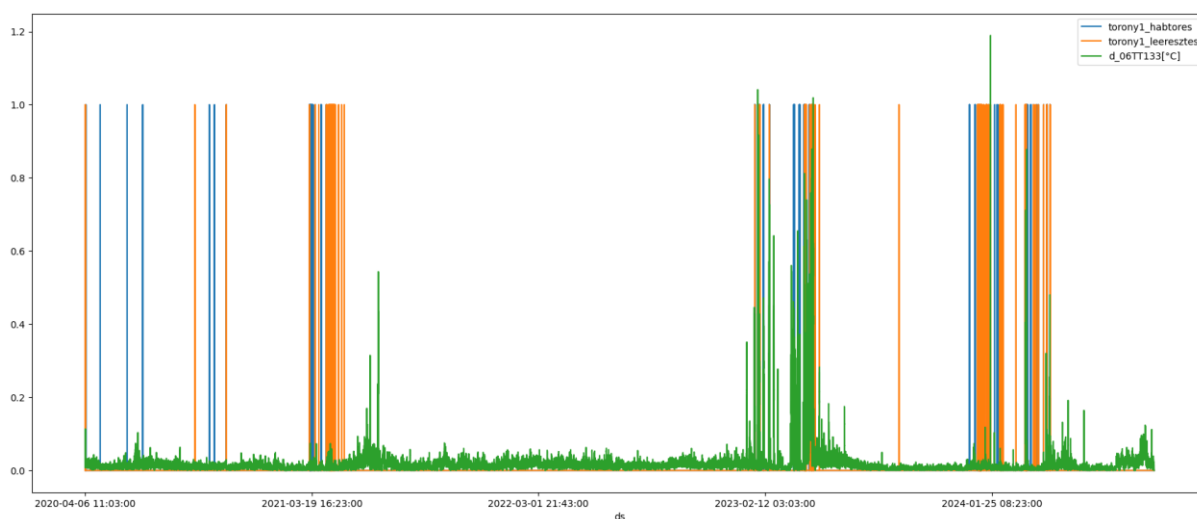
A nyers és az abszolút differencia értékek korrelációja az 1. műtárgy üzemjelére és méréseire

Az ábrán is látható, hogy míg a nyers jelek minimális, a differenciált jelek (a zaj mértékek) 2-3-szoros korrelációt mutatnak a habzásokra.

A legnagyobb mértékben korreláló jelek a következők:

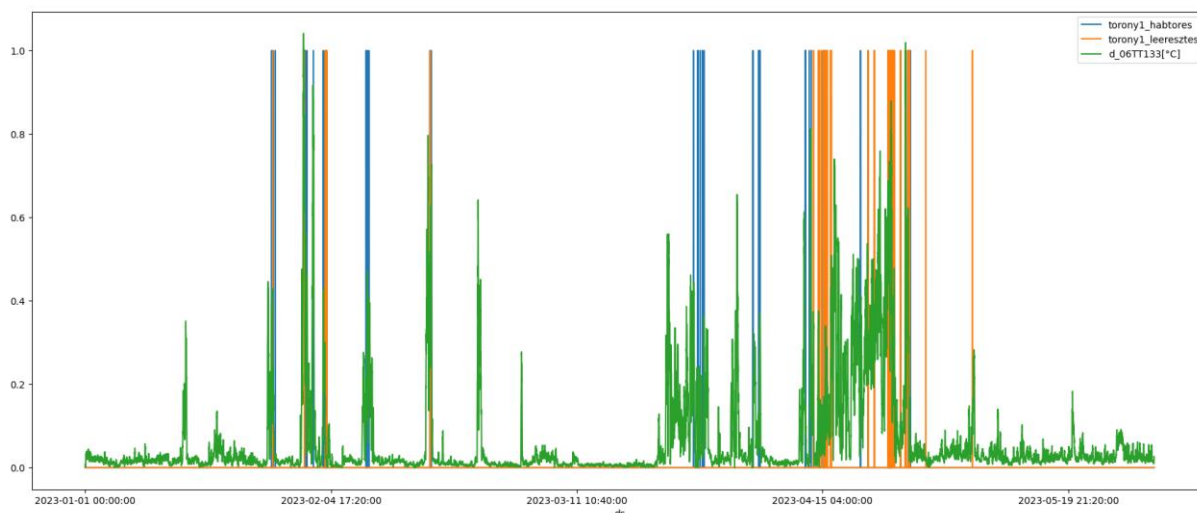
1. x12FTQ101BiogazMennyiségUgralas (*max: 0.2955*)
2. x06PT105BiogazNyomasUgralas (*max: 0.2625*)
3. x06TT133HomersekletUgralas (*max: 0.1480*)
4. x06TT141HomersekletUgralas (*max: 0.1031*)

A beavatkozások és a jelek közötti összefüggést érdemes idődiagramon is megvizsgálni:



Az 1. műtárgynál mért iszaphőmérséklet jelen vett zaj mértéke (zöld) és a beavatkozási üzemjelek (késsel a habtörés, narancs színnel az iszapleeresztés)

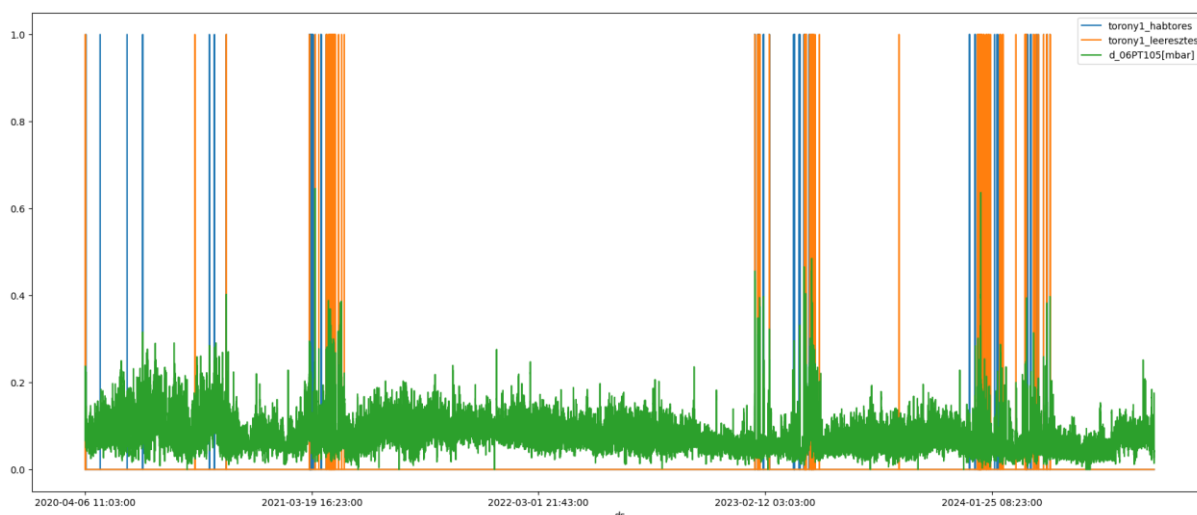
A diagramon látható, hogy a beavatkozások alatt, illetve körül a zaj mértéke nagyobb. Ebből még nem következtethetünk közvetlen ok-okozati viszonyra, de a kezdeti feltételezésünk részben megerősítést nyert. A pontosabb megértéshez érdemes a jelnek egy rövidebb időszakon vett értékét vizsgálni:



Az 1. műtárgyba beömlő iszap hőmérsékletén megfigyelt zaj mértéke a 2023. év első felében

A nagyított képen jobban látható, hogy a nagyobb mértékű zaj nemcsak a beavatkozásokat megelőzően, hanem azok során, illetve közvetlenül azokat követően is megmarad. Ez alapján azt az állítást tehetjük, hogy a nagyobb mértékű zaj nemcsak a habzást vagy más külső körülményt jelez, hanem utalhat az éppen folyamatban lévő beavatkozásra is.

Ezt az összefüggést ellenőrizhetjük másik jel diagramjának vizsgálatával is:



Az 1. műtárgyban mért relatív gáznyomás jelen tapasztalt zaj mértéke (zölddel) és a beavatkozások

A vizsgálat alapján kijelenthető, hogy a soft-szenzor tervezése során különös figyelmet kell fordítani az éppen zajló vagy nemrég lezajlott beavatkozások hatására, mint esetleges zavaró körülményre.

A zavaró hatásra tekintettel az osztályozó modell tanítását megelőzően az összes olyan adatpontot töröltem a tanító halmazból, ahol a műtárgyon beavatkozás zajlik, illetve rövid időn belül zajlott.

A jelek szemlélődő vizsgálatok azt is felfedezhetjük, hogy a zajnak, tehát a jel ugrálását okozó jelnek is jelentős „ugráló” természete van, azaz percről percre jelentős a megváltozása. Amennyiben ezt adatpontok megközelítésében értelmezzük, úgy a zajjal terhelt időszakban a rendszer állapota nem pusztán távolabbi adatpontokon tartózkodik, hanem azon túl is egymástól távoli adatpontok között ugrál. A gépi tanulás tervezésekor kézenfekvő lehet a rendszer állapotát egy szűkebb térrészre szorítani, hogy az osztályozás elvégezhető legyen. Ennek érdekében a jelen célszerű simítást végezni, aminek hatására a pillanatnyi állapotok helyett az adatpontok az ablakolt időszakban meglátogatott adatpontok tendenciózus elhelyezkedését kódolják. Ezzel az is elkerülhető, hogy a gyorsan változó értékek miatt a soft-szenzor pillanatról pillanatra változtatja a kimeneti jelét, felváltva habzást jelez és normális üzemi állapotot.

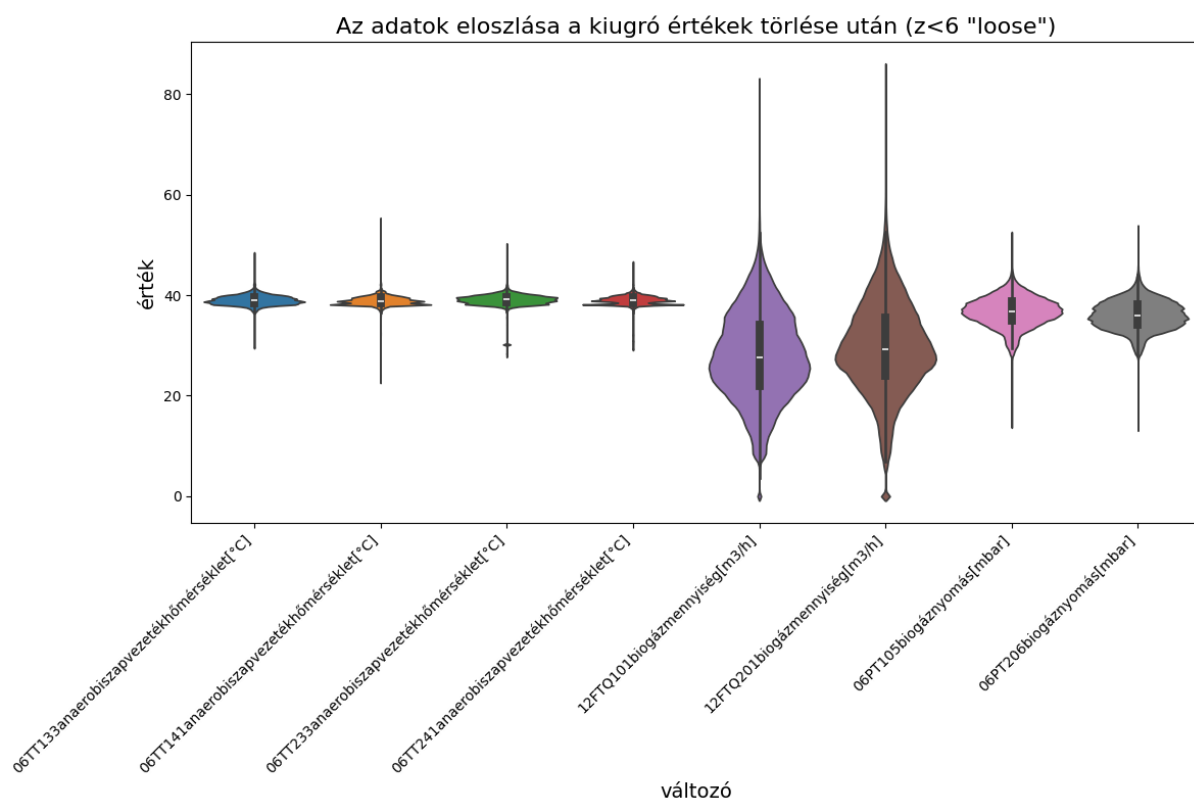
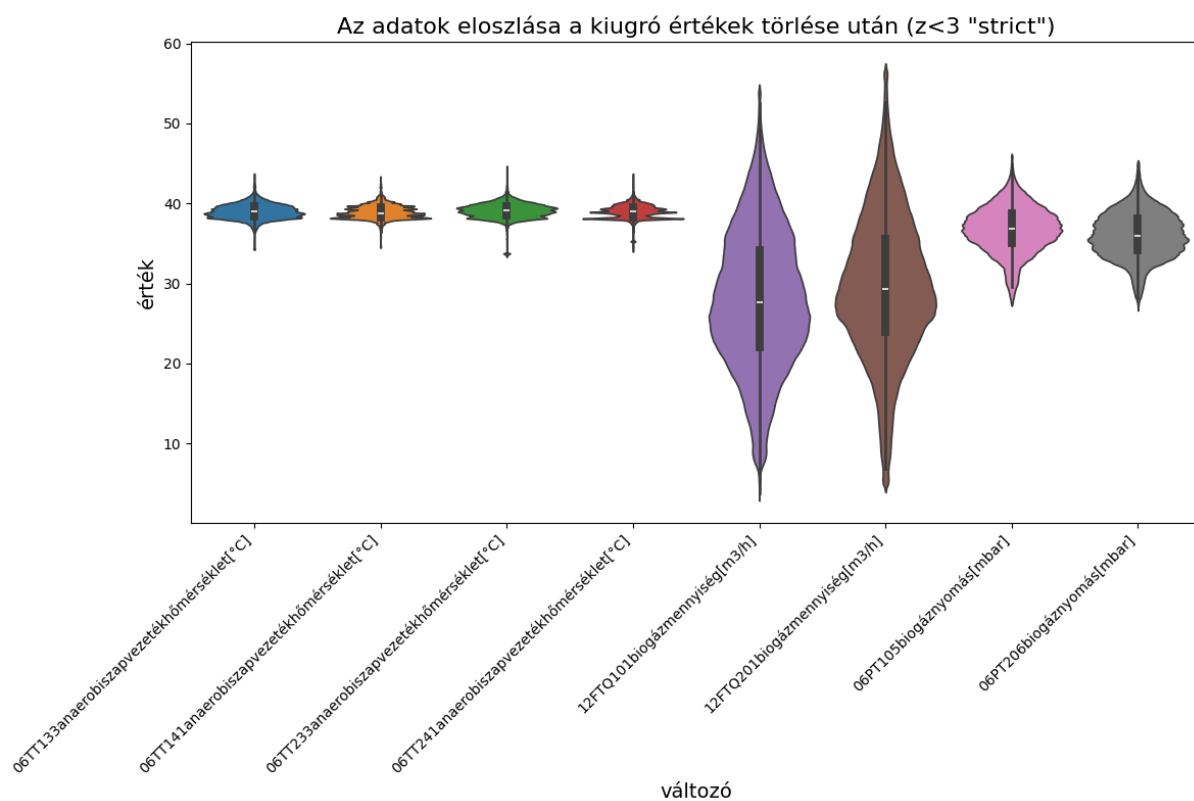
A simításhoz alkalmazott ablak méretét nem határoztam meg pontosan, de a jelek természete alapján a következő megállapításokat tehetjük: egyszerű mozgóátlag (SMA) alkalmazása esetén a simításnak legalább akkorának kell lennie, hogy eltűnjenek a jelből a pillanatszerű kiugrások. Az ablak nem lehet nagyobb, mint a beavatkozást megelőző maximális időintervallum kétszerese.

A simítás alkalmazásakor fontos további megkötéseket tenni. A simítás igazításakor szem előtt kell tartani, hogy a soft-szenzornak élő adatokon is működnie kell. Ezért tehát az ablakok csak lemaradó, azaz jobb oldalon számított átlagok lehetnek. Ez azonban elkerülhetetlenül a jelek késését eredményezi, mégpedig az ablak felének megfelelő idővel. A beavatkozásból származó zaj is nagyobb késéssel kódolódik az ablakba, így az abból fakadó zaj az ablak teljes szélességével nagyobb időben lesz hatással a jelre, az ablak szélességének felével a zaj megszűnte után még mindig a jelet 50%-ban meghatározza.

3.2.2 Adat előkészítése

Természetesen az adat részleges előkészítése elengedhetetlen volt a folyamat és a jelek megismeréséhez. Ezen túlmenően viszont, az adatok célzott átalakítását az Adat előkészítése rész alatt tárgyalom, mivel itt már az adatok gépi tanulással való feldolgozásához történnek átalakítások.

Az adatok egy nagyon nyers, közvetlenül a szenzorokról származó, tisztítatlan jelként érkeztek és kerültek archiválásra. A további használatához a kiugró értékek törlése mellett döntöttem. Tekintettel azonban arra, hogy a vizsgálat tárgya kifejezetten a zajjal terhelt, tehát hibás jelekre fókuszál, kétféleképpen töröltem a kiugró jeleket, és a továbbiakban ezeket külön adatsorként külön-külön is feldolgoztam. A szigorúbb tisztítás során a jeleknek egy felső korlátot szabtam (60 értékkel, ami kézenfekvő, hiszen a nyomás és a hőmérséklet is jóval alacsonyabb tartományban mozog), majd töröltem az átlagnál a szórás háromszorosánál nagyobb eltérésű értékeket ($|x-\mu|>3\sigma$). A másik megközelítésben nem vettem konstans felső korlátot az értékekre, illetve csak a szórás hatszorosánál nagyobb eltérésű jeleket töröltem ($|x-\mu|>6\sigma$).



A Kihívások részben tárgyaltam, hogy a jelekhez nem áll rendelkezésre közvetlen címkézett adat. A címkézést ezért a rendelkezésre álló jelek használatával kell megalkotni. A címkézett adatokat a következő feltevésekkel alkottam meg:

1. az iszapleeresztések nem állnak közvetlen összefüggésben a habzással, azonban zavarják a jelet a beavatkozás alatt és azt követően t_{NL} ideig,
2. a t_{NL} idő hossza függ az alkalmazott simító ablak t_{SM} szélességétől ($t_{NL} \gg t_{SM}/2$),
3. a habtörések a habzás közvetlen következményeként történnek,
4. a habtörést megelőző t_{FB} idő alatt a habzás következményként zavart jelek mérhetők,
5. a habzással zavart intervallum szélessége függ a simító ablak szélességétől ($t_{FB} \ll t_{SM}$)

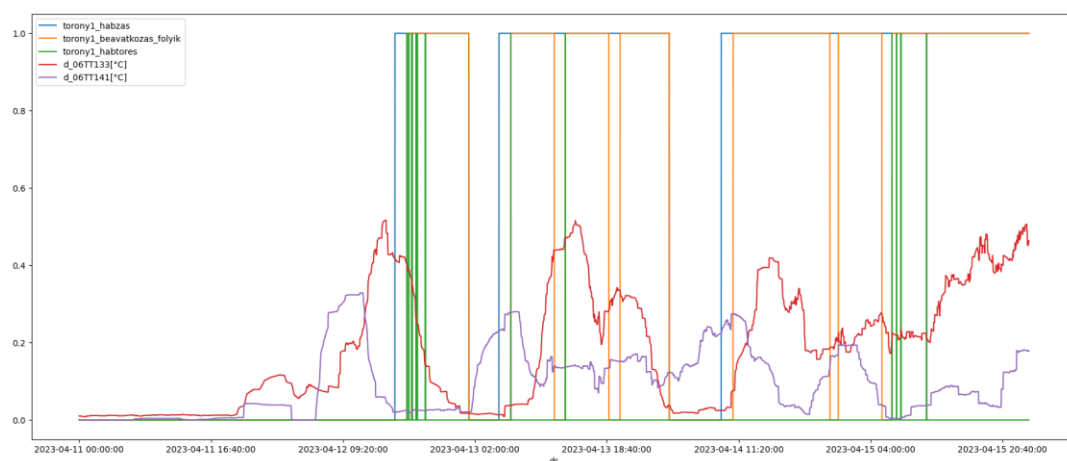
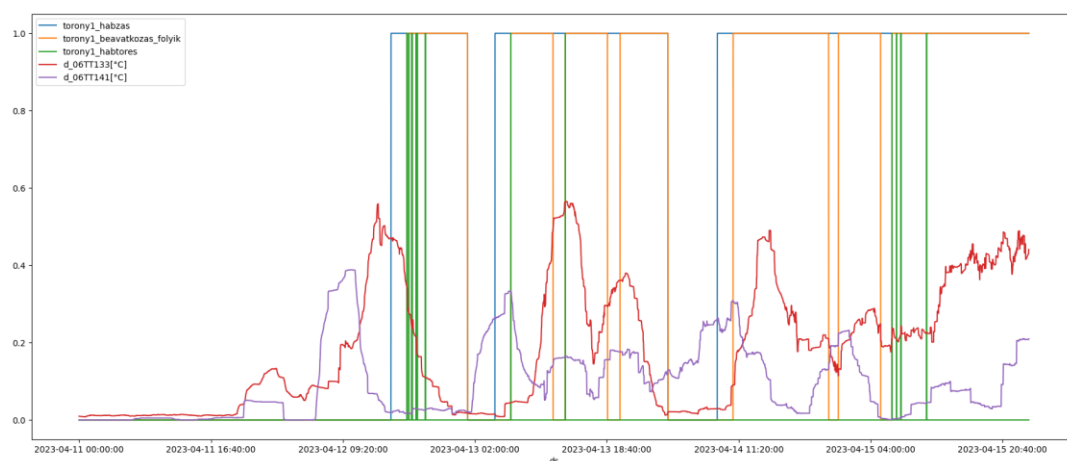
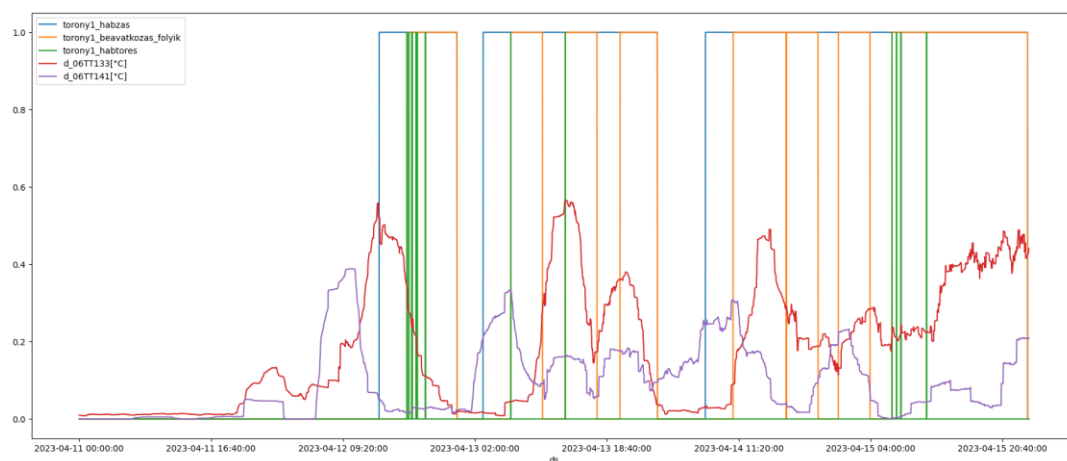
A feltevések nem pontosak és nem is helytállóak a rendszer gyakorlati működését tekintve, azonban lehetőséget teremtenek arra, hogy a szabályok alkalmazásával egy nem tökéletes címkézést hozhassunk létre az adatpontokon. Az osztályozó eljárás megalkotásakor ezért célszerű törekedni a kevésbé pontos, alultanultabb modellek megalkotására, így a rosszul címkézett adatpontok nem fogják súlyosan rontani az osztályozó modell általánosító képességét. A szabályrendszer a következő módon alkalmazható:

1. A szenzorokról származó jeleket egyazon t_{SM} szélességű lemaradó ablakkal simítjuk. Az ablak szélességét úgy választjuk meg, hogy a zavarjelek csúcsa a simítás hatására ne csússzon a beavatkozásra túlra, de a simítás megszüntesse a hirtelen kiugrásokat.
2. A habtörésekből és az iszapleeresztésekből közös beavatkozási jelet hozunk létre, amely jelet t_{NL} idővel nyújtunk. A t_{NL} megválasztásakor ügyelünk arra, hogy $t_{SM}/2$ -nél jelentősen nagyobb legyen, célszerűen úgy választjuk meg, hogy t_{SM} -nél is nagyobb legyen. Az így címkézett adatpontokat eldobjuk, mivel a gépi tanulás során nem választhatók le egyértelműen a habzással zavart illetve a normális üzemállapot jelekről. Ebből következik, hogy t_{NL} felső korlátjának megválasztásakor arra kell különösen ügyelni, hogy megfelelően nagy számú adatpont maradjon a gépi tanítás elvégzésére.
3. A habtörésekből származó üzemjelet az időben visszafelé terjesztjük t_{FB} intervallummal. A t_{FB} megválasztásakor arra ügyelünk, hogy jelentősen kisebb legyen, mint t_{SM} , célszerűen kisebb, mint $t_{SM}/2$. Az így címkézett adatok alkotják a *habzással* címkézett osztály tagjait, míg a megmaradó adatpontok lesznek a *nem habzás* címkével ellátott osztály elemei.

A bemutatott szabályrendszer természetesen több módon is teljesíthető. A kísérleteim során három különböző paraméterezést vizsgáltam meg. Ezek a következők:

1. $t_{SM} = 300$ (min); $t_{NL} = 240$; $t_{FB} = 210$
2. $t_{SM} = 300$ (min); $t_{NL} = 320$; $t_{FB} = 120$
3. $t_{SM} = 360$ (min); $t_{NL} = 330$; $t_{FB} = 90$

Az egyes beállítások hatása idődiagramon ábrázolva:



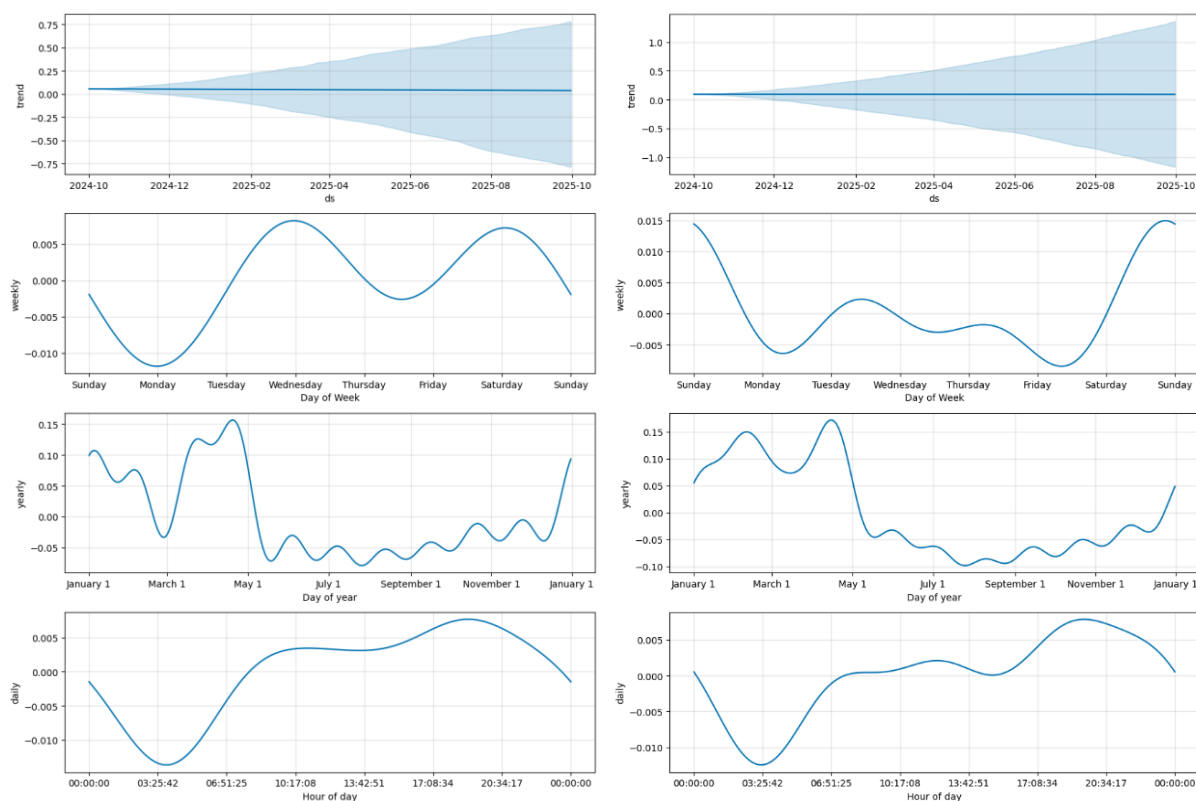
Három különböző paraméterezésű címkézési eljárás. Zölddel a habtörés beavatkozások pillanatszerűen, narancssal az eldobásra kerülő beavatkozástól zavart jelek, késsel a habzás címkével ellátott intervallum (Fentről lefelé: 1., 2. és 3. beállítás)

A kísérleteim tehát összesen 6 különbözően átalakított adathalmazon folytak, amelyekből 3 szigorúbb, 3 pedig lazább kiugróértékszabályok mentén került megalkotásra, és mindkét esetben az itt szemléltetett címkézési eljárásokat alkalmaztam.

3.2.3 További elemzés

A jelek időbeliségére tekintettel hasznosnak találtam, ha a modellek megépítése előtt a jeleket idősorelemzésnek is alávetem. Az elemzés célja, hogy a habzásról időbeli összefüggést is lehetőségem legyen megállapítani, illetve amennyiben egy erős modellt sikerül alkotni, úgy annak előrejelzését felhasználjam egy esetleges plusz feature-ként az osztályozó modellben. Ennek gyakorlati előnye lehetne, hogy ha nagy bizonyossággal sikerül megjósolni a közeljövőben tapasztalható zaj mértékét, illetve figyelembe véve a habzás bekövetkeztének szezonálisát, akár percekkel azelőtt adjon riasztást a rendszer, mint hogy a habzásra jellemző összetett állapot valójában megjelenne a mérési jeleken. Erre vonatkozó elképzeléseimet végül nem valósítottam meg, de az idősorelemzésből kiindulva így is hasznos következtetéseket tehetünk.

Az idősorelemzéshez a tisztított és címkézett adatsort használtam fel („laza” $z > 6$ tisztítású és $T_{SM}=300$; $T_{NL}=320$; $T_{FB}=120$), a korlátozottan rendelkezésre álló számítási kapacitás miatt alacsonyabb frekvenciával (óránként) újramintavételezve. A mintavételezés számtani átlag képzéssel történt. A dekompozíciót a Facebook Prophet modellel (Taylor-Letham, 2018) végeztem.

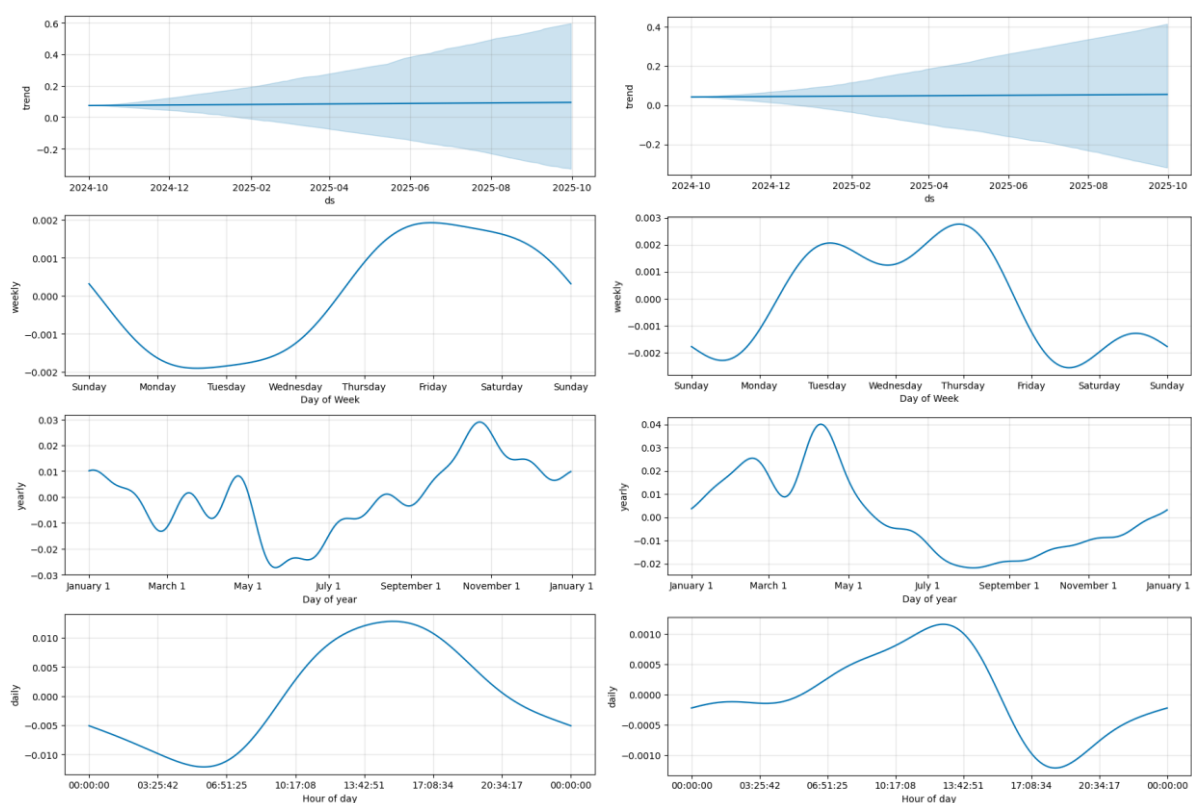


A habzási idősorok időkomponensek szerinti dekompozíciója. Balra az első, jobbra a második műtárgy

Az első műtárgyon tapasztalt habzási jel időkomponens dekompozíciójából látható, hogy a habzás valószínűsége tendenciáját tekintve állandó. Szezonálisát tekintve megállapítható, hogy legnagyobb mértékben az éven belüli szezonális befolyásolja a habzás gyakoriságát. Kiemelkedően magas a habzás valószínűsége a késő márciusi-áprilisi időszakban, maximumán 15%-kal gyakoribb a habzás bekövetkezése, mint átlagosan.

A dekompozíció ilyen módon történő elemzésekor azonban tekintetbe kell venni, hogy az éves szezonálisra vonatkozóan kevés adat áll rendelkezésre, csupán 3 teljes év. Ebből két évben valóban igen jellemző volt a tavasszal bekövetkező habzás, azonban a messzemenő következtetések levonása előtt mindenképp érdemes átgondolni, hogy van-e rendszerszerűen befolyásoló hatása ennek a magas kiugrásnak, vagy pusztán a rövid időhorizonttal magyarázható.

A második műtárgyon is elvégezhetjük a dekompozíciót. Ezen is látható, hogy az éven belüli szezonális a legjelentősebb, a csúcán (áprilisban) nagyjából 17%-kal magasabb a habzás bekövetkezésének gyakorisága, mint átlagosan. Ez nem jelenti következésképpen azt, hogy áprilisban a habzás valószínűsége 17%-kal nagyobb lenne, pusztán azt, hogy egy viszonylag rövid vizsgálati időszakban a gyakoriság ennyivel volt magasabb.



A zajok időbeliségének vizsgálata. Balra az első műtárgyon mért nyomás, jobbra a második műtárgyból mért gáztérfogatáram zajának mértéke

A dekompozíciót az egyes zaj jeleken külön is elvégezhetjük. Azt tapasztaljuk, hogy ilyenkor az éven belüli szezonális sokkal kisebb mértékben, csupán 3-4% körül jelenik meg a jelekben.

3.2.4 Modell építés

Az *Adat előkészítése* részben bemutatott 6 adathalmaz mindegyikét használtam a tanítások során, és a vizsgálat egyik tárgya volt, hogy melyik beállítás hozza a legjobb eredményeket egy ilyen speciális esetben, ahol a jeleken tapasztalt zaj adja a bemeneti értékeket az osztályozáshoz.

Az osztályozáshoz a következő osztályozó algoritmusokat alkalmaztam:

1. Logisztikus regresszió (*sklearn.linear_model.LogisticRegression*),
2. Support Vector Machine alapú lineáris osztályozó (*sklearn.svm.LinearSVC*)
3. K legközelebbi szomszéd osztályozó (*sklearn.neighbors.KNeighborsClassifier*),
4. Hisztogram alapú grádiens fokozó osztályozó
(*sklearn.ensemble.HistGradientBoostingClassifier*)
5. AdaBoost osztályozó (*sklearn.ensemble.AdaBoostClassifier*)
6. Többszörös döntési fa osztályozó (*sklearn.ensemble.RandomForestClassifier*)
7. Normál eloszlású naiv Bayes osztályozó (*sklearn.naive_bayes.GaussianNB*)

A logisztikus és Support Vector Machine alapú osztályozóknál a tanítások során a {0,25; 0,5; 1,0; 2,0; 4,0} halmazból válogattam paramétereiket.

A minta erős kiegyenlítetlenségére való tekintettel a tanítások felében alkalmaztam a SMOTEENN nevű alá- és fölé mintavételező kiegyenlítő eljárást (Batista et al., 2003)

3.2.5 Modell kiértékelés

Az osztályozó modellek kiértékelésére használtam a sztenderdnek tekinthető metrikákat (*accuracy*, *recall*, *precision*, *f1_score*). Ezek a metrikák azonban kifejezetten az osztályozó megközelítésre alkalmazhatók, ezért nem megfelelőek a létrejövő idősorok egymáshoz hasonlítására.

A vizsgálatom egyik célja, hogy az alkalmazott módszerrel elért beavatkozási minta összehasonlítható legyen más módszerekből származó beavatkozási mintákkal. Erre a célra leválasztottam a rendelkezésre álló minta egy összefüggő szakaszát, amit nem használtam fel a modell tanításakor, pusztán az értékelés során. Célszerűen a minta legfrissebb szakaszát vettem külön, ami a 2024 január 1. és szeptember 30. közötti intervallum, ami a teljes eredeti minta 17%-a.

Az időbeli jelek (tehát a 2024-es valós és predikált beavatkozási minta) hasonlóságának vizsgálatára ezért további metrikákat vezettem be. Az elsődleges ezek közül a *Fast Dynamic Time Warping* (Salvador-Chan, 2007) eljárással számolt jeltávolság. Ez nem azonos a (Sakoe-Chiba, 1978) eljárással, hanem egy nagyobb mintákra alkalmazható $O(n)$ osztályú algoritmussal számolt becslés a távolságra.

Az idősorok hasonlóságának vizsgálatára ezen kívül megszerkesztésre került a 60 perccel simított beavatkozójelek adatpontonként vett távolságának összege, ehhez hasonlóan a 60 perc szélességű ablakok maximumai között vett távolságok összege.

A modellek használhatóságában jelentős szerepet tölt be a jelzések gyakorisága a teljes idősorhoz, illetve az operátor beavatkozásainak számához viszonyítva. Ezekre is megalkotásra kerültek a megfelelő metrikák. A predikátumban található beavatkozások és az operátor beavatkozásainak számának különbsége illetve hányadosa is naplózásra került a modellekkel.

Dur	Name	c	classif	outliers	T _{SM}	T _{NL}	T _{FB}	SMOTEENN	t	acc	prec	recall	f1↑
11.2s	sedate-slug		naiveb	loose	300	320	120	False	1	0.955	0.076	0.306	0.122
11.4s	whimsical-shoat		naiveb	loose	360	330	90	False	1	0.959	0.055	0.314	0.093
1.2min	shivering-snake	.25	svm	loose	360	330	90	True	1	0.955	0.020	0.462	0.039
1.2min	suave-yak	4	svm	loose	360	330	90	True	1	0.955	0.020	0.462	0.039
13.8s	luminous-horse		naiveb	loose	300	320	120	False	2	0.947	0.016	0.117	0.029
11.9s	abrasive-kit		naiveb	loose	360	330	90	False	2	0.953	0.013	0.131	0.024
1.2min	overjoyed-asp		naiveb	loose	300	320	120	True	2	0.915	0.010	0.172	0.020
1.1min	dapper-seal		naiveb	loose	360	330	90	True	2	0.917	0.007	0.168	0.013

Dur	Name	c	classif	dtw	mean_diff	max_diff	T _{op_foam}	T _{pred_foam}	T _{add_foam}	T _{mply_foam}
11.2s	sedate-slug		naiveb	6619	15259	18858	4330	12885	8555	2.9758
11.4s	whimsical-shoat		naiveb	3552	13862	16992	3222	12243	9021	3.7998
1.2min	shivering-snake	.25	svm	7068	15127	21156	3222	12795	9573	3.9711
1.2min	suave-yak	4	svm	7068	15130	21142	3222	12799	9577	3.9724
13.8s	luminous-horse		naiveb	4980	17829	21995	8025	12841	4816	1.6001
11.9s	abrasive-kit		naiveb	3791	15611	19410	6005	12019	6014	2.0015
1.2min	overjoyed-asp		naiveb	9434	28655	34139	8025	24485	16460	3.0511
1.1min	dapper-seal		naiveb	8664	27654	32499	6005	25022	19017	4.1669

A legjobbnak ítélt tanított osztályozó modellek. c az SVM és logisztikus osztályozók c paramétere, t-vel rövidítve a műtárgy sorszáma

A metrikákon alkalmazott szűrések:

DTW távolság max 10000;

T_{mply}(predikció beavatkozási idő az operátoré arányában) min. 0.2 max 8

rendezés: f1 score szerint csökkenő

félkövérrel kiemelve: 7000 alatti DTW távolság

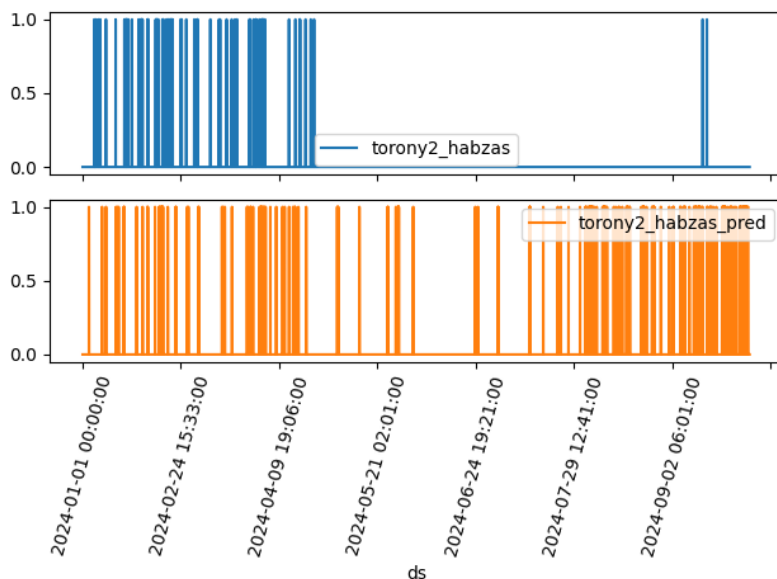
A kísérletben létrejött modellek vizsgálatokor feltűnően jó eredményt értek el a Gaussian Naive Bayes alapú osztályozók. A legjobban illeszkedő modellek összehasonlításakor feltűnhet, hogy a Naive Bayes osztályozók rendre jobb eredményt értek el a kiegyensúlyozatlan adathalmazon. Ez az állítás egyedül erre a módszerre alkalmazható, a többi osztályozó nem volt képes értékelhető eredményt adni a kiegyensúlyozatlan halmazon.

Az összehasonlítás céljára alkotott metrikák kifejezetten hasznosnak bizonyultak bizonyos extrém esetek kiszűrésére. Az osztályozók egy része a 2024-es különálló mintára szélsőséges jelet alkotott, ami üres, vagy majdnem üres idősort (minden adatpont *nem habzás* alá sorolása) vagy túlzásfolt idősort (szinte minden adatpont *habzás* alá sorolása) eredményezett. A hagyományos metrikák ezekben az esetekben azért nem voltak kellően jól használhatók, mert az összehasonlításokor használt idősorok is erősen kiegyenlítettlen mintát tartalmaztak.

A legjobbnak ítélt modellek metrikáit vizsgálva feltűnhet, hogy valójában alacsony értékeket kaptunk. A legmagasabb *f1 score* is csupán 12%, ami rendkívül alacsonynak mondható. A teljes kísérleti halmazon a legmagasabb *f1 score* valójában 98%, de abban az esetben a predikált habzásban telt percek száma mindösszesen 1 volt. Ebből a példából is látható, hogy az erősen kiegyensúlyozatlan mintából fakadóan a hagyományos metrikák nem jól alkalmazhatók erre az adathalmazra. Az alábbi összevetés jól szemlélteti, hogy a hagyományos metrikák milyen módon torzítják a kapott eredmények hitelességét:

Dur	Name	c	classif	outliers	T _{SM}	T _{NL}	T _{FB}	SMOTEENN	t	acc	prec	recall	f1
2.2min	lyrical-moth-52		knn	loose	360	330	90	True	2	0.9532	0.9365	0.9840	0.9597

Az itt bemutatott K-nearest neighbor modell a hagyományos metrikák szerint igen jó eredményt ért el a 2024-es adatsoron alkalmazva. 95.9%-os *f1 score*-ja alapján egy jól alkalmazható modellnek gondolhatjuk



Ugyanennek a modellnek a 2024-es idősoron mutatott viselkedése, idődiagramon ábrázolva

Jól látható, hogy bár a két idősor mutat hasonlóságokat, azonban a predikátumban jelentős számú habzás osztályú adatpont található, ami a valódi idősoron nem található meg. A kiegyenlítetlenség olyan jelentős, hogy valójában bármely modell, ami sok valós negatív értéket talál, elég jó metrikákat érhet el a kiértékeléskor.

A modell gyengeségeit jobban kiszűrhetjük, ha megvizsgáljuk az erre a célra hozzáadott metrikákat.

Name	dtw	mean_diff	max_diff	T _{op_foam}	T _{pred_foam}	T _{add_foam}	T _{mply_foam}
lyrical-moth-52	11137	15793.5	30695	6005	10675	4670	1.778

Új mérőszámok az eredmény kiértékeléséhez

A *FastDTW*-vel kapott távolság 11137, ami közepesnek mondható (az értékek átlaga 16230, 14012 szórással). A simított jelek távolsága 15793.5, ami kifejezetten jónak mondható (átlag: 27099; szórás: 19653), a maximumokkal ablakolt jelek különbsége 30695 (átlag: 35757; szórás: 26092). Azt láthatjuk tehát a jelek hasonlóságának vizsgálatakor, hogy a vizsgált modell csak kicsivel ad jobb eredményt az átlagosnál, csupán az osztályok kiegyenlítetlensége okozza a kiemelkedően jó metrikákat.

Ha a szűrt modellek közül választunk egyet alaposabb vizsgálatra, akkor az egyik kiemelkedő modell lehet az alábbi:

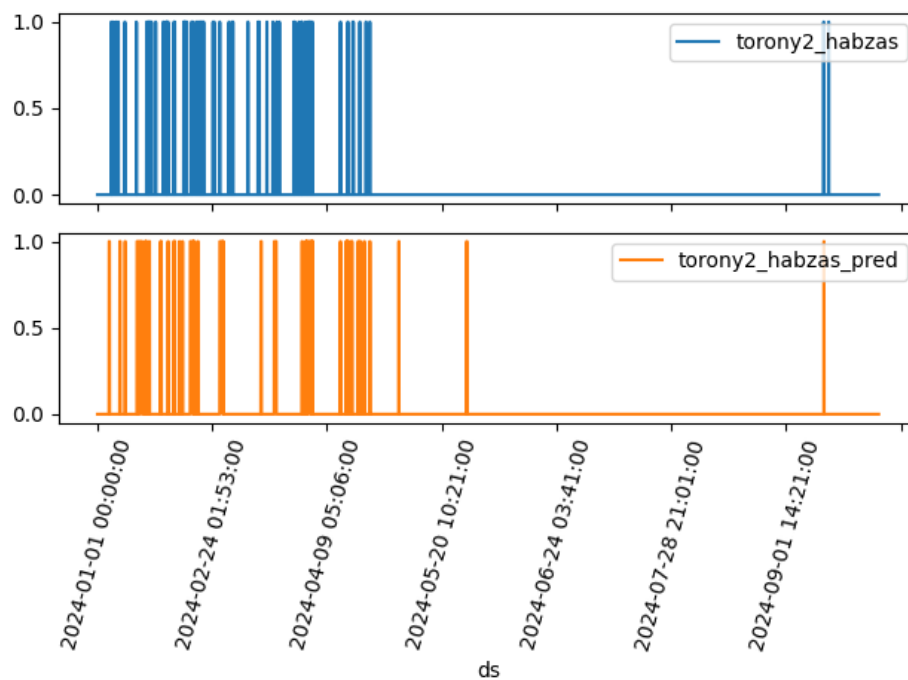
Dur	Name	classif	outliers	T _{SM}	T _{NL}	T _{FB}	SMOTEENN	t	acc	prec	recall	f1
13.8s	luminous-horse	naiveb	loose	300	320	120	False	2	0.95	0.0164	0.1166	0.0288

dtw	mean_diff	max_diff	T _{op_foam}	T _{pred_foam}	T _{add_foam}	T _{mply_foam}
4980.0000	17828.53	21995	8025	12841	4816	1.60012

A FastDTW távolság és a maximumokból képzett jeltávolság jóval kisebb mint az előző példában

Ebben az esetben a becsült DTW jeltávolság korlátozásával éltünk. Az osztályozó metrikák ebben az esetben lényegesen alacsonyabbak, az *f1 score* mindössze 2,9%, ami értelmezhetetlenül alacsonynak tűnik.

Hasonlítsuk össze a 2024-es időszakra alkotott mintát idődiagramon!



A Naive Bayes osztályozó predikátuma a 2024-es intervallumon

A különbség szembetűnő. A habzás predikátumok sűrűsége ott koncentrálódik, ahol az operátor beavatkozásait megelőző címkézett adatpontok találhatók. Csupán két rövid időszak adódik a köztes időszakban, ahol a modell habzást predikált, azonban a beavatkozások szerint azon a részen nem történt igazolható habzás.

Itt fontos újra leszögezni, hogy a valódi állapotok utólagos ellenőrzésére nincs lehetőség, azonban a soft-szenzor tervezésekor érdemes szem előtt tartani, hogy túl sok riasztás esetén (még ha a riasztások megalapozottak is) a szenzor elveszítheti a használhatóságát, mivel az operátor figyelmét túlterheli. Ez alapján feltételezhető, hogy egy ritkábban esetleg pontatlanabbul, de az operátor viselkedéséhez hasonlóbban működő tanított modell hasznosabbnak bizonyulhat, mint egy olyan, ami minden potenciális habzási állapotra riasztást ad.

További kézenfekvő megoldás lehet a gyakrabban és a ritkábban jelző szenzorokból összetett (*ensemble*) modellt alkotni, amely esetleg súlyozott átlaggal számolt figyelmeztetési indexet ad, amelynek értelmezése, illetve riasztássá való átalakítása már az operátor felelőssége.

A futások összegző jellegű értékelésekor megállapíthatjuk, hogy az osztályozó eljárások közül a legsikeresebb (a gyakorlati alkalmazhatóságot is figyelembe véve) a Gaussian Naive Bayes klasszifikáló volt. Ezt követte a lineáris Support Vector Machine alapú LinearSVC osztályozó, azt pedig a Logisztikus regresszió eljárás.

Az adathalmazok közül az eredményesebb modellek között legnagyobb gyakorisággal a „lazán” tisztított (csak a 6 feletti z-score értékű változóktól megtisztított) adatsor volt. Azon belül is a $T_{SM}=360$; $T_{NL}=330$; $T_{FB}=90$ paraméterezésű konfiguráció.

A hozzáadott metrikák közül a *FastDTW*-vel becsült jeltávolság, a maximummal való ablakozásból vett távolság és a predikált illetve valós kitüntetett állapotok aránya (a szakirodalomban a PPV reciproka) bizonyult kiemelkedően hasznosnak. A *FastDTW* becslés helyett pontosabb lett volna a *DTW* eljárást alkalmazni, ez azonban $O(n^2)$ osztályú, ami ekkora mintára már nem alkalmazható.

4 Összegzés

4.1 Értékelés

Munkám során egy gyakorlatból vett, kevés mérési szemponttal bíró adathalmazon hajtottam végre átalakításokat és a jeleket terhelő zaj alapján osztályozó modelleket alkottam a habzás nevű kitüntetett állapot felismerésére. A tanított modelleket a hagyományos osztályozási metrikákkal és továbbiakkal összehasonlító jelleggel kiértékeltem. A *FastDTW* távolság metrikaként való alkalmazásával egy hatékonyan kiértékelő mértéket alkottam meg a predikátumként előálló és a gyakorlati rendszerből vett idősorok szemléletes összehasonlításához. Az alkalmazott megkötésekkel az ígéretesnek tűnő modellek közül ki tudtam választani azokat, amik egy gyakorlatban megvalósított soft-szenzor számára hasznos osztályozó jeleket nyújthatnak.

4.2 További megfontolások

A bemutatott munkát gyakorlati, célvezérelt fejlesztési feladatként értelmezve az alábbi következtetések tehetők:

1. A tanításhoz a beavatkozásokról megbízhatóbb adathalmazra van szükség, nem üzemidő jelekből előállítottra. A soft-szenzor futtatásához mindenképp érdemes frissíteni a rendelkezésre álló jeleket, hogy az átalakítások ne okozzanak értékbeli hibát. (Ebben a munkában az üzemjelek megalkotásához visszafele terjesztést alkalmaztam, amire egy élő rendszeren nincs lehetőség, előreterjesztett üzemjelekkel egyperces lemaradások jöhetnek létre a kikapcsolások során.)
2. A technológiai rendszernek csak bizonyos elemeiről sikerült adatot nyerni. A kapcsolódó kutatásokból egyértelmű, hogy a habzás szorosan összefügg az online mérőkkel nyomon követhető kémiai állapotokkal. Jelen munkám különlegessége, hogy ezek használata nélkül alkottam meg az osztályozó modellt, de feltételezhető, hogy a gyakorlatban pontosabb becslések alkothatók az anyag más, megfelelő állapotváltozóinak figyelembevételével (pl. szennyvíziszap pH kémhatása, gázkeverék CO₂ koncentrációja)
3. Az implementációban szereplőnél korábbi beavatkozás kialakítható akkor, ha megfelelő ablakolást választunk. Az itt bemutatott implementációban három különböző paraméterezést vizsgáltam meg, mindegyiket egyszerű mozgóátlag (SMA) számítással. Érdemes lehet megvizsgálni olyan modelleket, amit súlyozott (WMA) vagy exponenciális (EMA) ablakolással simítunk, így hamarabb kaphatunk kimagasló értékeket, amik szükséges beavatkozást jelezhetnek. WMA simítást alkalmazott (Kazemi et al. 2021)
4. A kísérletezés során csak egyszerűbb, könnyen tanítható osztályozók közül válogattam. Az eredményeket figyelembe véve a Support Vector Machine ígéretes módszer lehet egy hatékony osztályozó megalkotásához. Ezt a kérdést érdemes megvizsgálni alaposabban, szélesebb körben paraméterezett osztályozók tanításával, illetve

különböző alakú kernelekkel – a kísérlet során kizárólag lineáris kernellel végeztem tanítást.

5. A habzási események gyakoriságának figyelembevételével a Facebook Prophet időszorelemzés alkalmas lehet akár a kiugró értékek keresésére is. Ezek a kiugró értékek jelezhetnek habzást is, így ezt a kérdéskört is érdemes lehet megvizsgálni alaposabban. A témában csupán az elmúlt pár évben is több kutatás született. (Thiyagarajan et al. 2020), (Nithish et al. 2021), (Oswal et al. 2022)

4.3 Köszönetnyilvánítás

Elsősorban köszönettel tartozom Dr. Dulai Tibor tanár úrnak a témavezetésért és a szakmai irányításért.

Köszönöm Dr. Berzsenyi Miklós segítségét a kutatási téma megtalálásához és a közös munka megalapozását a résztvevőkkel. Köszönöm Kertai Zoltán úr hozzájárulását az adatok kinyeréséhez, Molnár Tamás úr segítségét az iszapkezelési technológiai információk terén.

Köszönöm a Pannon Egyetem Műszaki Informatikai Karának a 2024/2025 I. félévi Kari Tudományos Diákköri Ösztöndíjat, aminek támogatásával létrejött ez a kutatás.

4.4 Hivatkozások

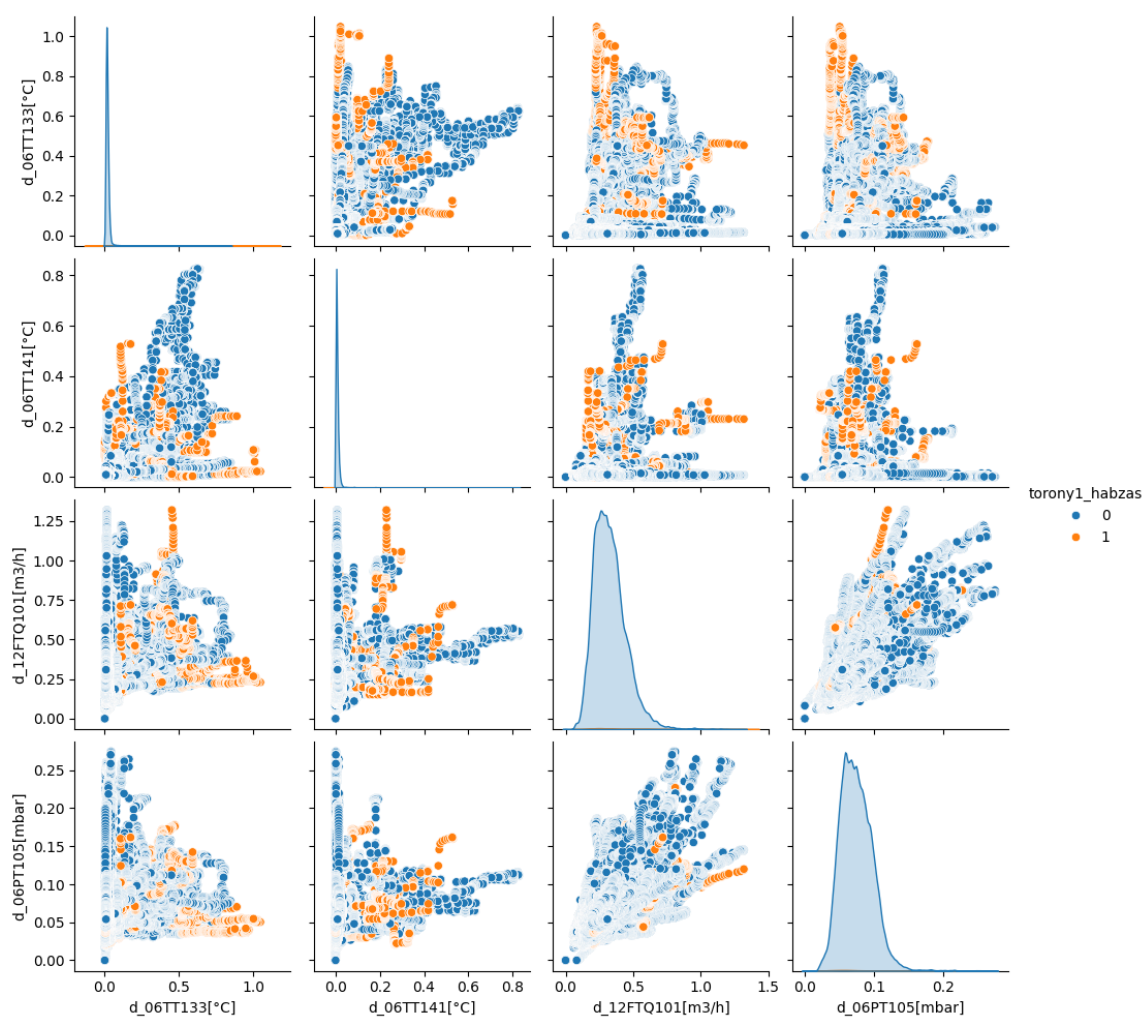
1. Kazemi, Pezhman et al. "Data-driven techniques for fault detection in anaerobic digestion process." *Process Safety and Environmental Protection* 146 (2021): 905-915.
2. Steyer, Jean-Philippe, et al. "Hybrid fuzzy neural network for diagnosis-application to the anaerobic treatment of wine distillery wastewater in a fluidized bed reactor." *Water Science and Technology* 36.6-7 (1997): 209-217.
3. Taylor, Sean J., and Benjamin Letham. "Forecasting at scale." *The American Statistician* 72.1 (2018): 37-45.
4. Studer, Stefan, et al. "Towards CRISP-ML (Q): a machine learning process model with quality assurance methodology." *Machine learning and knowledge extraction* 3.2 (2021): 392-413.
5. Batista, Gustavo EAPA, Ana LC Bazzan, and Maria Carolina Monard. "Balancing training data for automated annotation of keywords: a case study." *Wob* 3 (2003): 10-18.
6. Stan Salvador and Philip Chan. 2007. Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal.* 11, 5 (October 2007), 561–580.
7. Thiyagarajan, Karthick, et al. "A temporal forecasting driven approach using facebook's prophet method for anomaly detection in sewer air temperature sensor system." 2020 15th IEEE Conference on industrial electronics and applications (ICIEA). IEEE, 2020.
8. Nithish, T., et al. "Real-time anomaly detection using Facebook prophet." *International Journal of Natural Computing Research (IJNCR)* 10.3 (2021): 29-40.
9. Oswal, Sangeeta, Shubham Hadawle, and Atharva Khangar. "Anomaly Detection in Time Series Data by Forecasting Using Facebook Prophet." *International Advanced Computing Conference*. Cham: Springer Nature Switzerland, 2022.

5 Adatok és vizualizációk

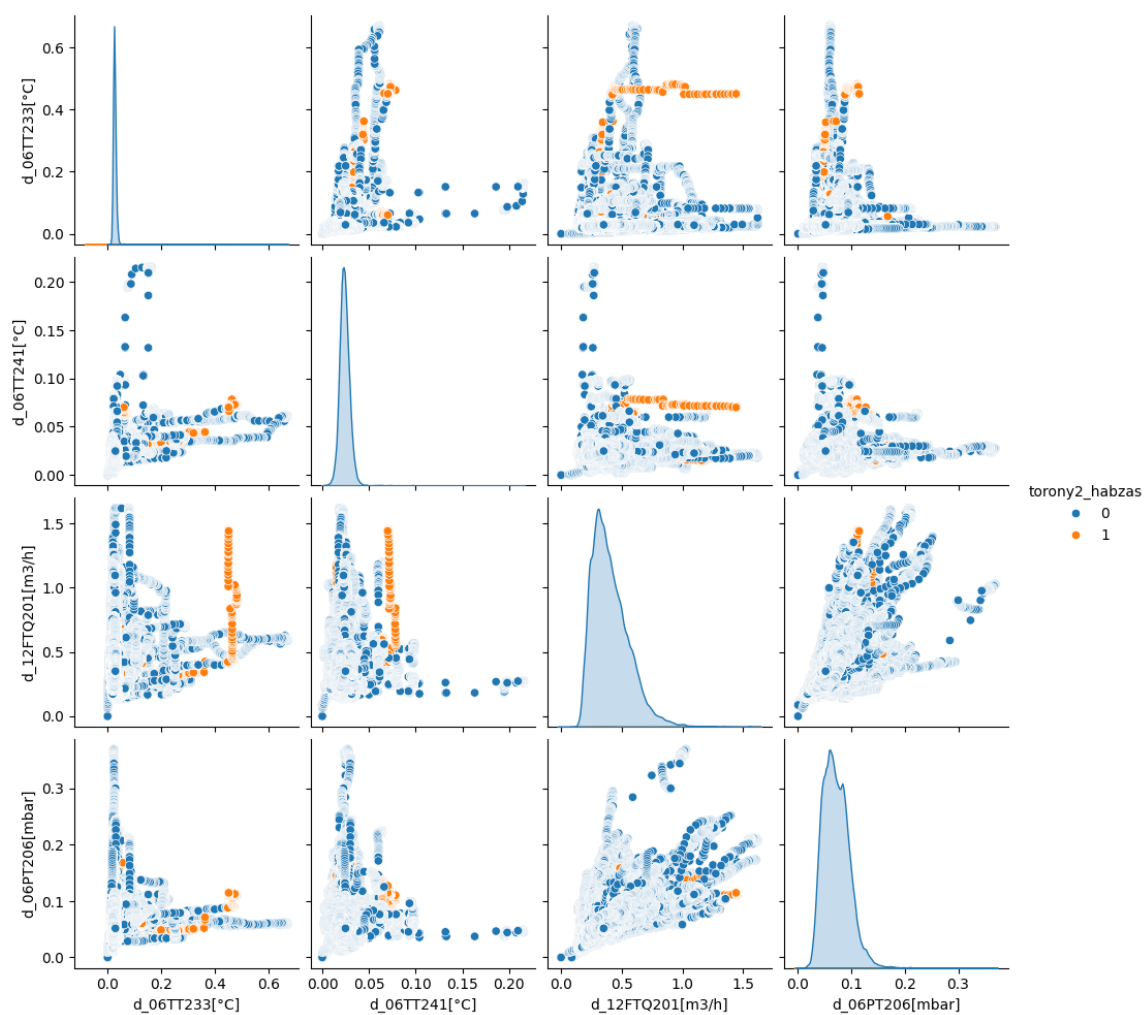
5.1 Páros eloszlások

5.1.1 Kiegyenlítés nélkül

első műtárgy:

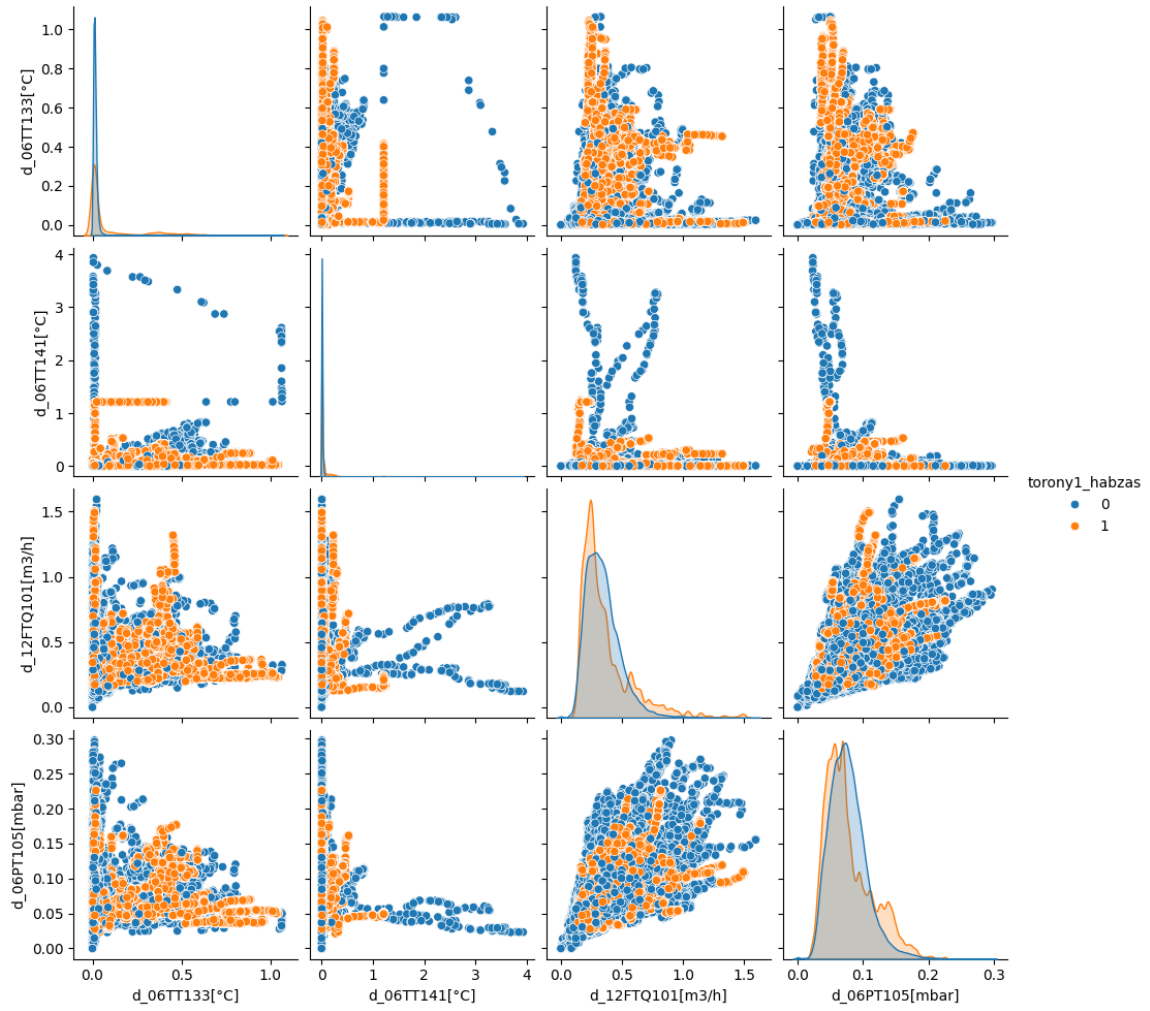


második műtárgy:

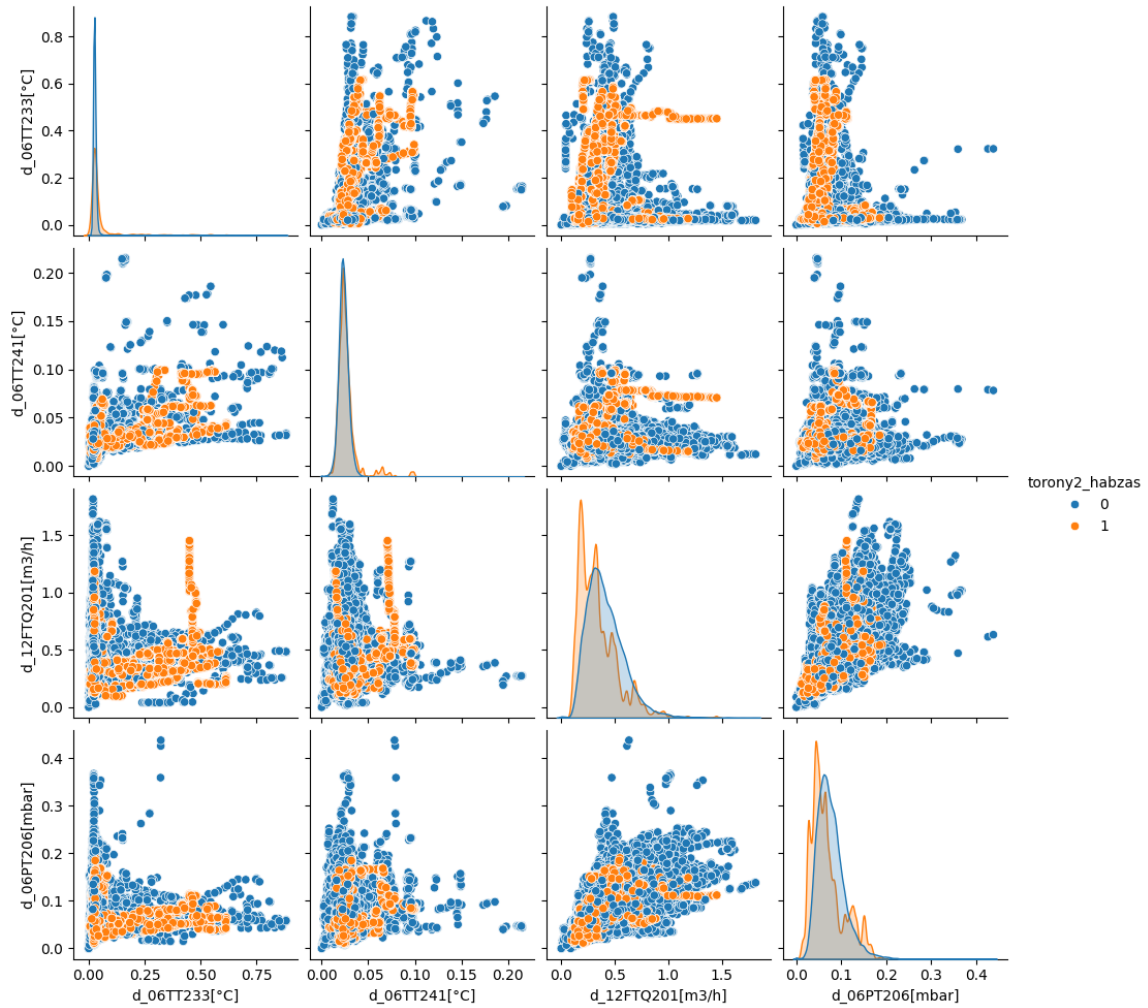


5.1.2 SMOTEENN kiegyenlítéssel

első műtárgy:



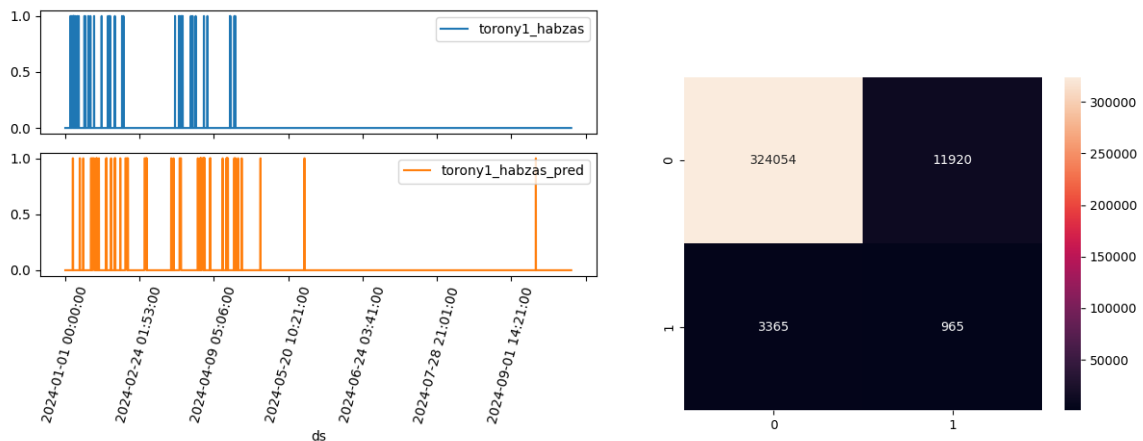
második műtárgy:



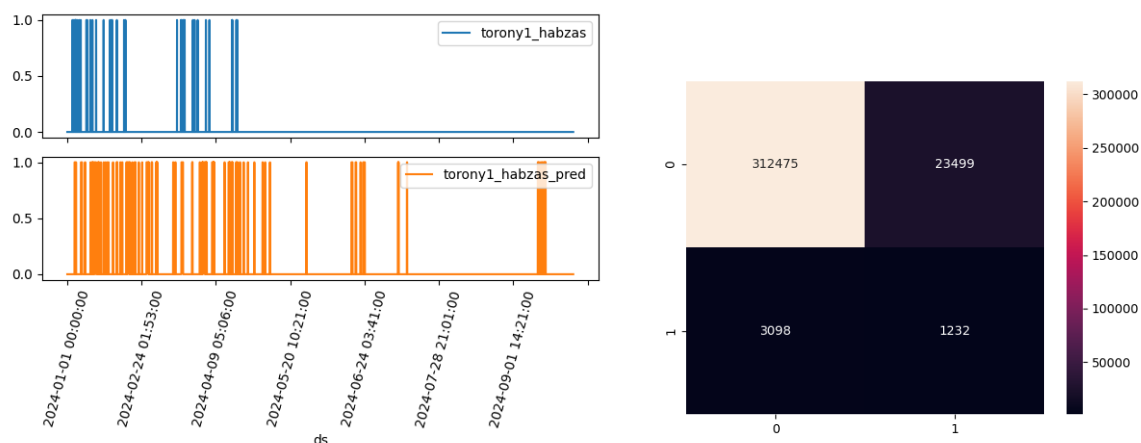
5.2 Idősor összehasonlítások és konfúziós mátrixok

5.2.1 Gaussian Naive Bayes osztályozó modell

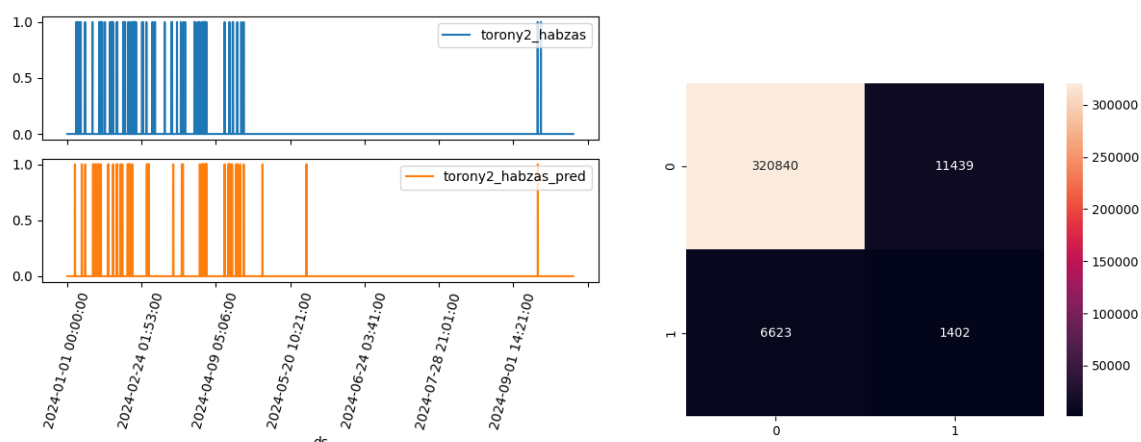
„laza” kiugróérték szabály alkalmazva; $T_{SM}=300$; $T_{NL}=320$; $T_{FB}=120$; első műtárgy; nem történt előzetes kiegyenlítés:



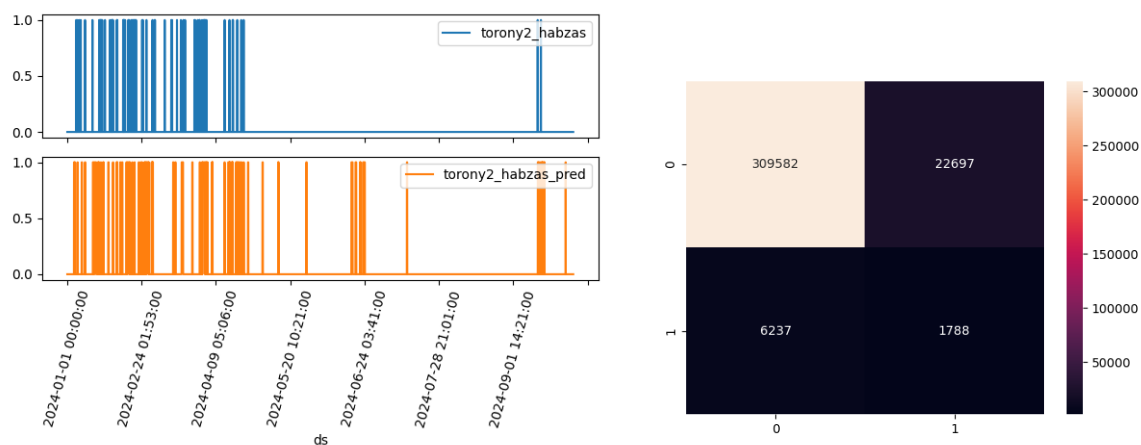
„laza” kiugróérték szabály alkalmazva; $T_{SM}=300$; $T_{NL}=320$; $T_{FB}=120$; első műtárgy; SMOTEENN előzetes kiegyenlítés:



„laza” kiugróérték szabály alkalmazva; $T_{SM}=300$; $T_{NL}=320$; $T_{FB}=120$; második műtárgy; nem történt előzetes kiegyenlítés:

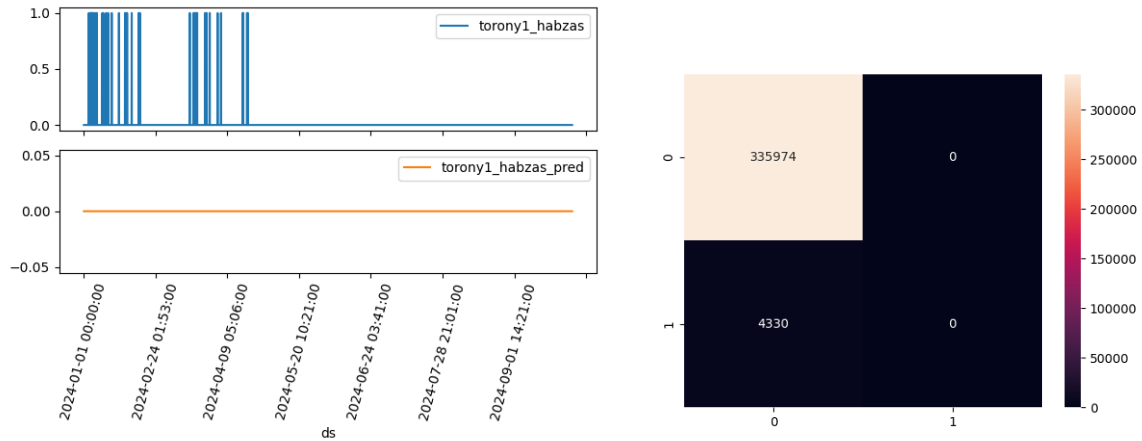


„laza” kiugróérték szabály alkalmazva; $T_{SM}=300$; $T_{NL}=320$; $T_{FB}=120$; második műtárgy; SMOTEENN előzetes kiegyenlítés:

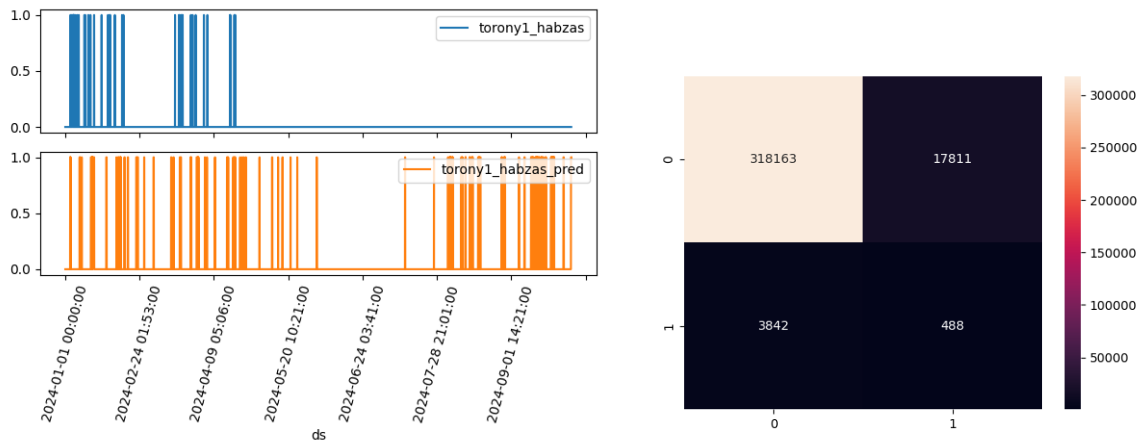


5.2.2 LinearSVC osztályozó modell

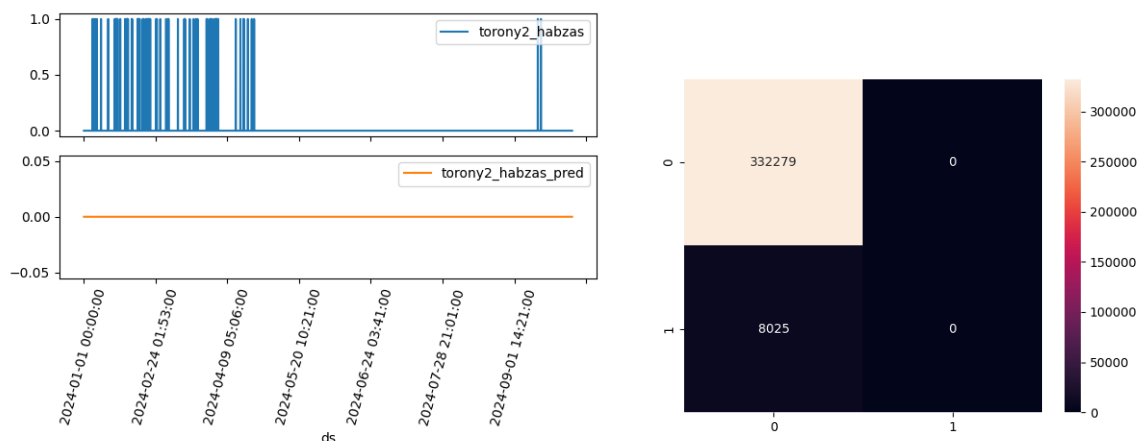
„laza” kiugróérték szabály alkalmazva; $T_{SM}=300$; $T_{NL}=320$; $T_{FB}=120$; első műtárgy; nem történt előzetes kiegyenlítés:



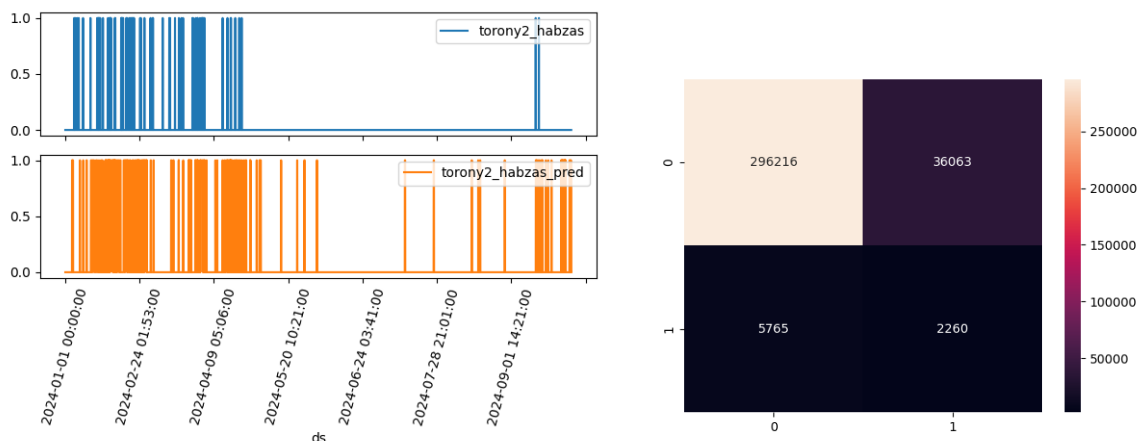
„laza” kiugróérték szabály alkalmazva; $T_{SM}=300$; $T_{NL}=320$; $T_{FB}=120$; első műtárgy; SMOTEEN előzetes kiegyenlítés:



„laza” kiugróérték szabály alkalmazva; $T_{SM}=300$; $T_{NL}=320$; $T_{FB}=120$; második műtárgy; nem történt előzetes kiegyenlítés:

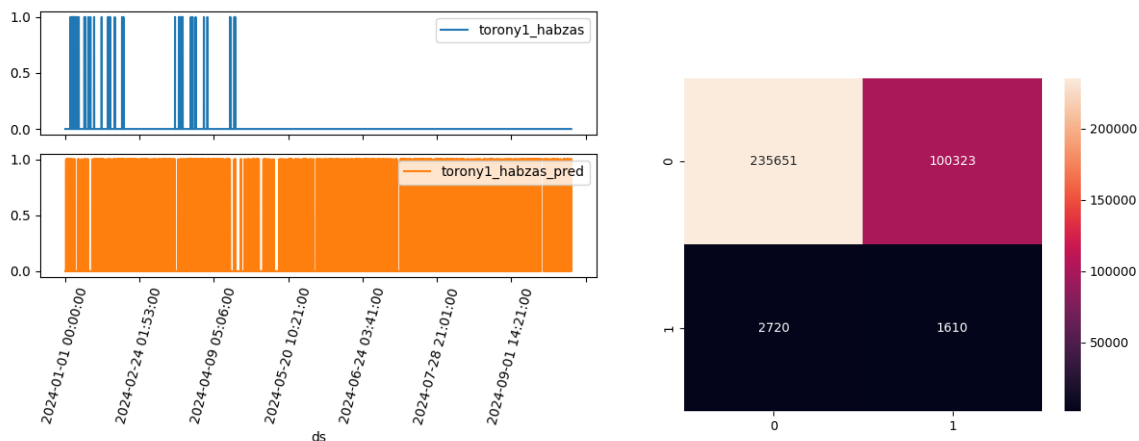


„laza” kiugróérték szabály alkalmazva; $T_{SM}=300$; $T_{NL}=320$; $T_{FB}=120$; második műtárgy;
SMOTEENN előzetes kiegyenlítés:

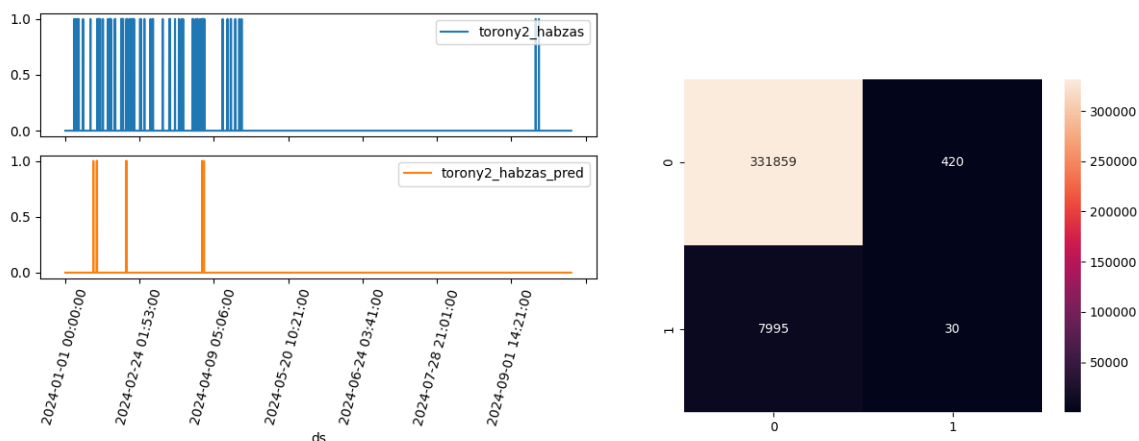


5.2.3 AdaBoost osztályozó

„laza” kiugróérték szabály alkalmazva; $T_{SM}=300$; $T_{NL}=320$; $T_{FB}=120$; első műtárgy;
SMOTEENN előzetes kiegyenlítés:

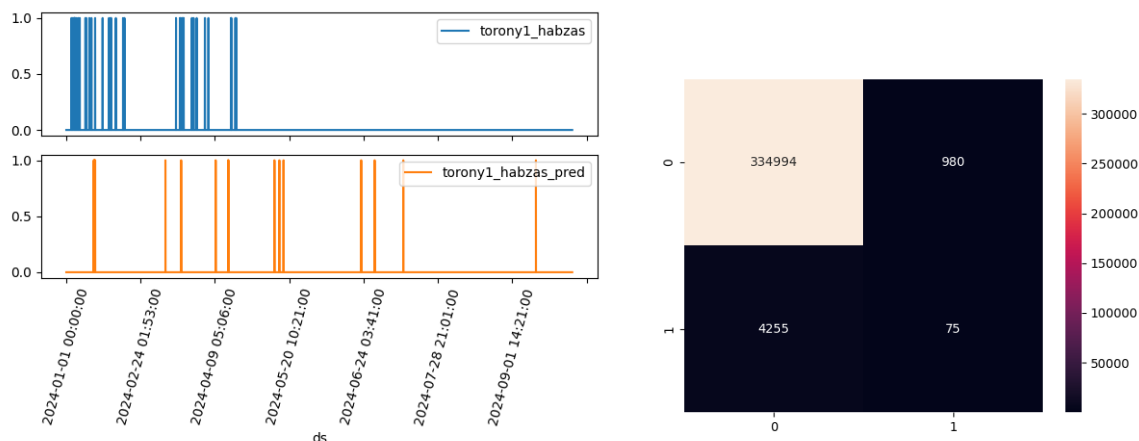


„laza” kiugróérték szabály alkalmazva; $T_{SM}=300$; $T_{NL}=320$; $T_{FB}=120$; második műtárgy;
nem történt előzetes kiegyenlítés:

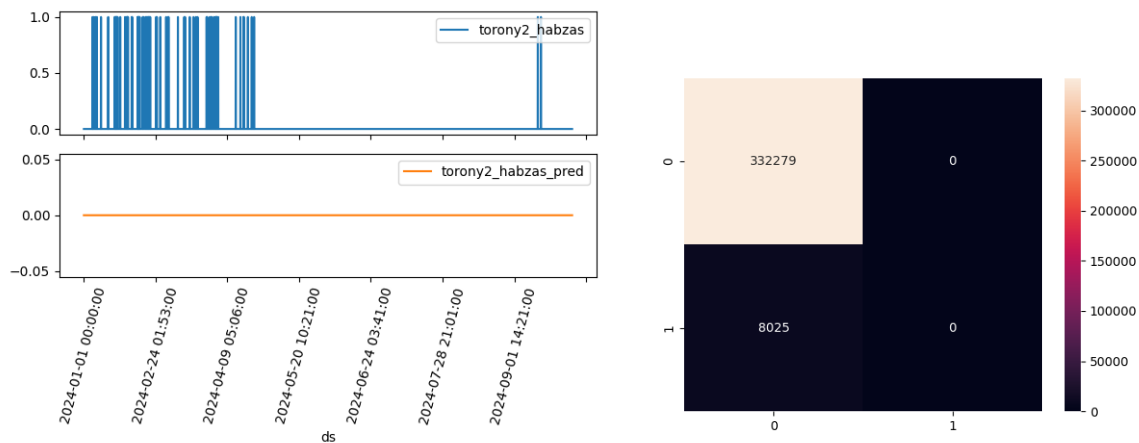


5.2.4 RandomForest osztályozó

„laza” kiugróérték szabály alkalmazva; $T_{SM}=300$; $T_{NL}=320$; $T_{FB}=120$; első műtárgy; nem történt előzetes kiegyenlítés:

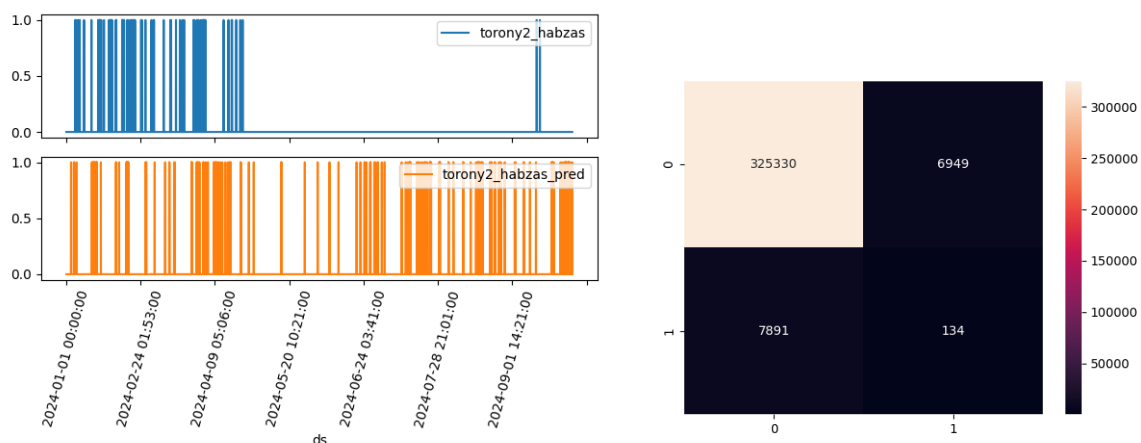


„laza” kiugróérték szabály alkalmazva; $T_{SM}=300$; $T_{NL}=320$; $T_{FB}=120$; második műtárgy; SMOTEENN előzetes kiegyenlítés:



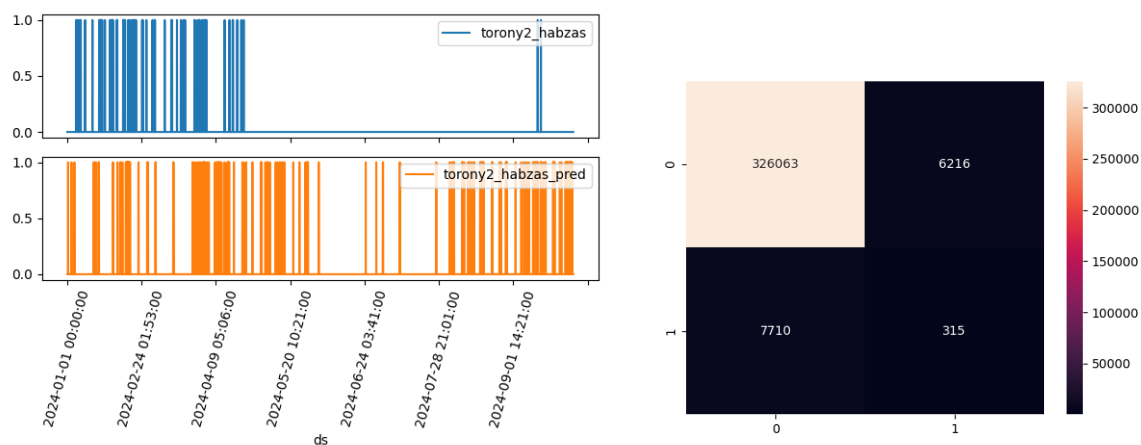
5.2.5 HistGradientBoosting osztályozó

„laza” kiugróérték szabály alkalmazva; $T_{SM}=300$; $T_{NL}=320$; $T_{FB}=120$; második műtárgy;
SMOTEENN előzetes kiegyenlítés:

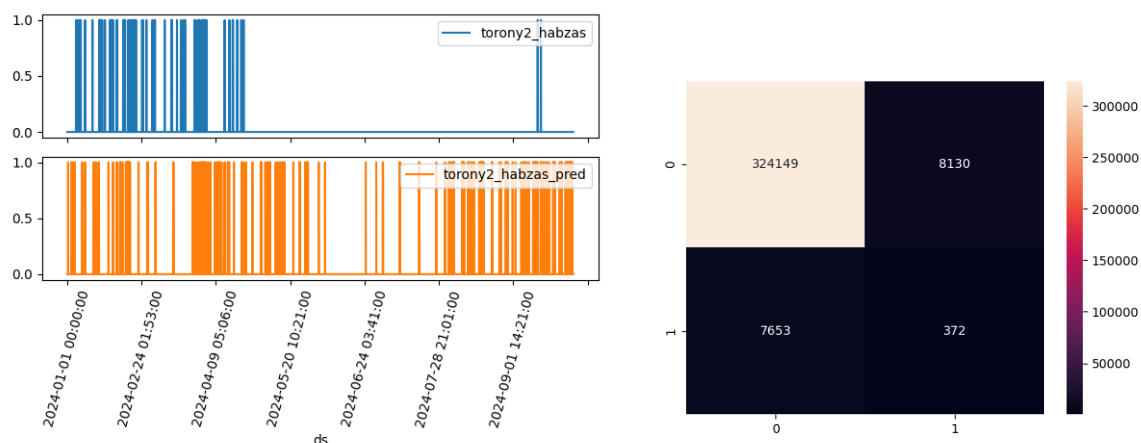


5.2.6 KNeighbors osztályozó

„laza” kiugróérték szabály alkalmazva; $T_{SM}=300$; $T_{NL}=320$; $T_{FB}=120$; második műtárgy;
nem történt előzetes kiegyenlítés:

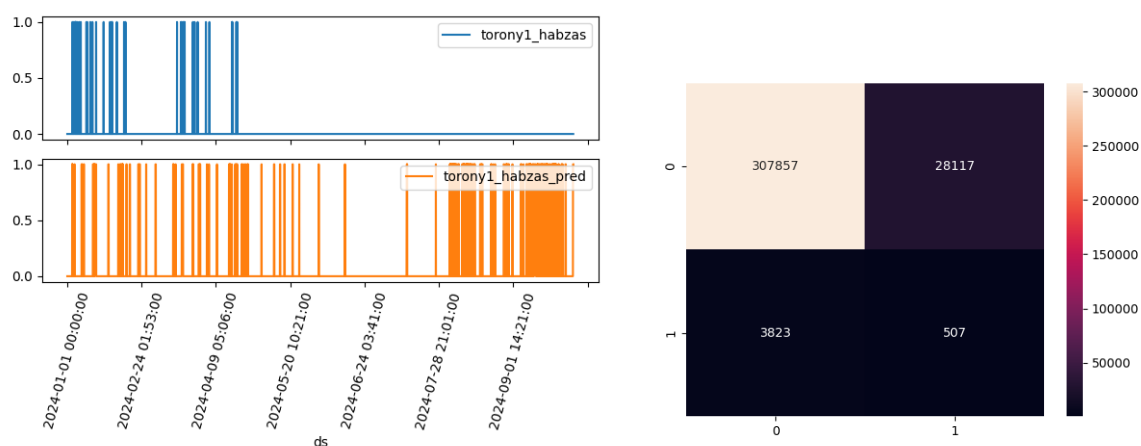


„laza” kiugróérték szabály alkalmazva; $T_{SM}=300$; $T_{NL}=320$; $T_{FB}=120$; második műtárgy; SMOTEENN előzetes kiegyenlítés:



5.2.7 LogisticRegression osztályozó

„laza” kiugróérték szabály alkalmazva; $T_{SM}=300$; $T_{NL}=320$; $T_{FB}=120$; első műtárgy; SMOTEENN előzetes kiegyenlítés:



5.3 Táblázatok

5.3.1 A jobbnak ítélt tanított modellek paraméterei és metrikái

Dur.	Name	c	classifier	dataset	SMOTEENN	tor.	accuracy	acc_2024	precision	prec_2024	recall	rec_2024	f1	f1_2024	dtw	mean_diff	max_diff	T _{op_foam}	T _{pred_foam}	T _{add_foam}	T _{mply_foam}
2.2min	lyrical-moth-52		knn	loose_df_3_sm_360_nl_330_fb_90	True	2	0.9998	0.9532	0.937	0.937	0.984	0.984	0.9597	0.9597	11137	15794	30695	6005	10675	4670	1.7777
11.2s	sedate-slug-705		naiveb	loose_df_3_sm_300_nl_320_fb_120	False	1	0.9761	0.9551	0.076	0.076	0.306	0.306	0.1217	0.1217	6619	15259	18858	4330	12885	8555	2.9758
12.8s	thoughtful-roo-299		naiveb	loose_df_3_sm_300_nl_320_fb_120	False	1	0.9761	0.9551	0.076	0.076	0.306	0.306	0.1217	0.1217	6619	15259	18858	4330	12885	8555	2.9758
11.4s	whimsical-shoat-202		naiveb	loose_df_3_sm_360_nl_330_fb_90	False	1	0.9760	0.9590	0.055	0.055	0.314	0.314	0.0933	0.0933	3552	13862	16992	3222	12243	9021	3.7998
1.2min	shivering-snake-347	0.25	svm	loose_df_3_sm_360_nl_330_fb_90	True	1	0.9100	0.9545	0.020	0.020	0.462	0.462	0.0388	0.0388	7068	15127	21156	3222	12795	9573	3.9711
1.2min	suave-yak-669	4	svm	loose_df_3_sm_360_nl_330_fb_90	True	1	0.9098	0.9545	0.020	0.020	0.462	0.462	0.0387	0.0387	7068	15130	21142	3222	12799	9577	3.9724
13.8s	luminous-horse-511		naiveb	loose_df_3_sm_300_nl_320_fb_120	False	2	0.9785	0.9469	0.016	0.016	0.117	0.117	0.0288	0.0288	4980	17829	21995	8025	12841	4816	1.6001
11.9s	abrasive-kit-306		naiveb	loose_df_3_sm_360_nl_330_fb_90	False	2	0.9787	0.9533	0.013	0.013	0.131	0.131	0.0236	0.0236	3791	15611	19410	6005	12019	6014	2.0015
1.2min	overjoyed-asp-980		naiveb	loose_df_3_sm_300_nl_320_fb_120	True	2	0.9527	0.9150	0.010	0.010	0.172	0.172	0.0195	0.0195	9434	28655	34139	8025	24485	16460	3.0511
12.4s	puzzled-snail-811	4	log	loose_df_3_sm_360_nl_330_fb_90	False	1	0.9961	0.9895	0.438	0.438	0.009	0.009	0.0183	0.0183	3072	3555	5644	3222	354	-2868	0.1099
12.3s	melodic-bear-628	2	log	loose_df_3_sm_360_nl_330_fb_90	False	1	0.9961	0.9895	0.438	0.438	0.009	0.009	0.0183	0.0183	3072	3555	5644	3222	354	-2868	0.1099
12.2s	languid-rook-54	0.5	log	loose_df_3_sm_360_nl_330_fb_90	False	1	0.9961	0.9895	0.438	0.438	0.009	0.009	0.0183	0.0183	3072	3556	5645	3222	355	-2867	0.1102
1.1min	clean-dog-981		naiveb	strict_df_3_sm_300_nl_240_fb_210	True	2	0.9206	0.8363	0.008	0.008	0.216	0.216	0.0157	0.0157	11152	58589	63968	11144	58403	47259	5.2408
1.1min	dapper-seal-222		naiveb	loose_df_3_sm_360_nl_330_fb_90	True	2	0.9486	0.9173	0.007	0.007	0.168	0.168	0.0127	0.0127	8664	27654	32499	6005	25022	19017	4.1669
13.3s	brawny-horse-420	0.5	log	loose_df_3_sm_360_nl_330_fb_90	False	2	0.9980	0.9795	0.000	0.000	0.000	0.000	0.0000	0.0000	5760	6953	10960	6005	948	-5057	0.1579
13.2s	amusing-jay-476	1	log	loose_df_3_sm_360_nl_330_fb_90	False	2	0.9980	0.9792	0.000	0.000	0.000	0.000	0.0000	0.0000	5760	7058	11065	6005	1053	-4952	0.1754