BÁNKI KÖZLEMÉNYEK --- BÁNKI REPORTS
VI. ÉVFOLYAM 2. SZÁM --- VOLUME VI NO. 2

Kibertámadások előre jelzése historikus adat alapú előre jelző rendszerrel és Fuzzy logika alapú kockázatbecsléssel

Prediction of cyber attacks with historical data-based forecasting system and Fuzzy logic-based risk estimation

Krasnyánszki Brúnó*, Prof. Dr. Zlatko Čović**, Prof. Dr. Rajnai Zoltán***
* Óbudai Egyetem Bánki Donát Gépész és Biztonságtechnikai Mérnöki Kar, Budapest, Magyarország
** Szabadkai Műszaki Szakfőiskola, Szabadka, Szerbia
*** Óbudai Egyetem Bánki Donát Gépész és Biztonságtechnikai Mérnöki Kar, Budapest, Magyarország
brunokrasnyanszki@stud.uni-obuda.hu, chole@vts.su.ac.rs , rajnai.zoltan@bgk.uni-obuda.hu

***Összefoglalás* — Kutatásunkban az EuRepoC adatbázisát használtuk a jövőbeli kibertámadások előrejelzésére. Neurális hálózatokat, támogató vektorgép modelleket és Fine Tree algoritmust használtunk. Emellett egy Fuzzy alapú kockázatelemzőt is fejlesztünk az államilag támogatott kibertámadásokhoz. Kutatásunk folytatódik, így csak részeredmények vannak a cikkben.**

***Kulcsszavak*:** EuRepoC, Kibertámadások előrejelzése, predikció, AI, MI

***Abstract* — In our research paper we used the EuRepoC's database to make predictions for the future cyberattacks. We used neural networks, support vector machine modells and Fine Tree algorythm. We are also developing a Fuzzy based risk analyzer for state sponsored cyber attacks. Our research is continuing so there are only partial results in the paper.**

***Keywords*:** EuRepoC, Prediction of cyber attacks, prediction, AI

1 BEVEZETÉS

1.1 Motiváció

A kutatás ihletét az adta, hogy az idei félévben az Erasmus+ program során Prof. Dr. Zlatko Čović által részt vehettünk egy cég látogatáson a Studio Present cégnél Szabadkán. Ahol bemutatták, hogy ők fejlesztették a European Repository of Cyber Incidents (EuRepoC) weboldalát. Ezen weboldalon rengeteg korábbi támadást dokumentáltak és gyűjtöttek össze. Prof. Dr. Zlatko Čović-al való beszélgetésünk során azon kezdtünk el gondolkodni, hogy mit lehetne ezekkel az adatokkal kezdeni.

1.2 Kutatás hasznosulásának lehetőségei

Amennyiben kutatásunk eredményei hosszabb távon pozitív eredményt hoznak abban az esetben a korább mintázatok adatai alapján képesek lehetünk bizonyos napokat vagy időszakokat kockázatosabbnak prediktálni ezáltal előre tudunk készülni humán erőforrássokkal és egyéb intézkedésekkel

2 EUREPOC

„A European Repository of Cyber Incidents (EuRepoC) egy független kutatókonzorcium, amelynek célja a kiberfenyegetések környezetének jobb megértése az Európai Unióban és azon kívül. A 2022 novemberében elindított fő célunk az adatközpontú megbeszélések és politikaalkotás előmozdítása a kiberbiztonság területén, valamint a kiberbiztonsági fenyegetések tudatosítása. Ezt úgy érjük el, hogy elemzési keretet biztosítunk a kiberincidensek „életciklusának" értékeléséhez és összehasonlításához, a technikai, politikai és jogi szempontokra összpontosítva. Adataink és kutatásaink az érintettek széles köre számára relevánsak – beleértve a kormányzati tisztviselőket, a civil társadalom képviselőit, az üzleti élet vezetőit, az újságírókat, az oktatókat, a diákokat és a nagyközönséget." [1]

2.1 EuRepoC tevékenysége

2.1.1 Adatbázis készítése a kiberincidensekről

Az EuRepoC napi rendszerességgel dokumentálja a nyilvános incidenseket, valamint elemzéseket készít, amit elérhetővé tesz adatbázisában a weboldalukon.[1]

2.1.2 Kutatások és elemzések készítése

Az EuRepoc a kiberkonfliktusok trendjeit empirikus módszerekkel elemzi és akadémiai igényességgel készít cikkeket. Napi szinten is készít rövid jelentéseket, valamint konferenciákon megosztják az aktualitásokat a kiberincidenserőL.[1]

2.1.3 Európai hálózat építése

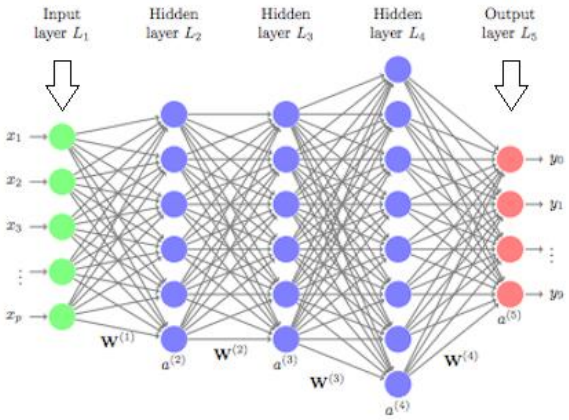
Speciális oktatási programokkal és tudományos kerekasztal-beszélgetésekkel erősítik a független kiberbiztonsági kutatószervezetek és kutatók európai hálózatát. [1]

3 TECHNOLÓGIAI BEVEZETŐ

Az alábbiakban kívánjuk ismertetni azokat a technológiákat, amelyeket felhasználtunk a kutatásunk során.

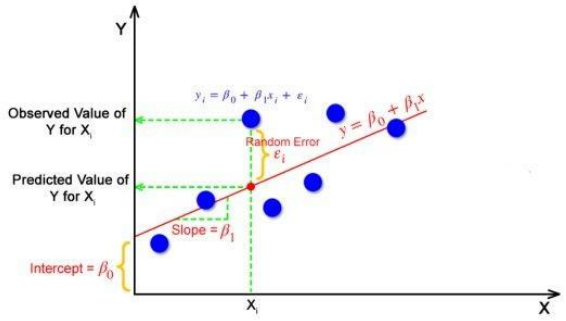
3.1 Gépi taunlás

A gépi tanulás az olyan algoritmusok és statisztikai modellek fejlesztésére összpontosító mesterséges intelligencia (AI/MI) egyik alágaként definiálható, amelyek lehetővé teszik a számítógépek számára, hogy konkrét feladatokat végezzenek el kifejezett utasítások nélkül. Ehelyett ezek a rendszerek a tapasztalatokból tanulnak és javulnak, az adatokban rejlő minták elemzésével és azonosításával. [2]



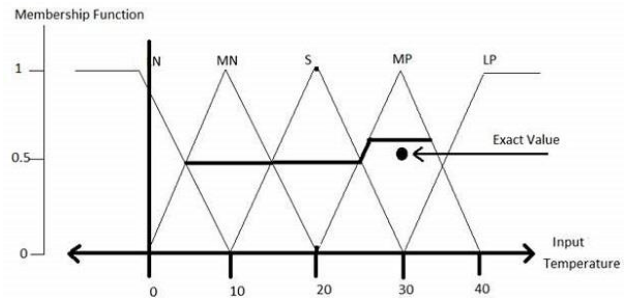
3.2 Predikció

"A predikció a gépi tanulásban egy olyan folyamat, amely során a modell a bemeneti adatok alapján megpróbálja előrejelezni a kimeneti értékeket. Ez a modell tanulási fázisában szerzett tapasztalatokra és a bemeneti adatokban található mintákra épül."[2]



3.3 Fuzzy Logika

"A fuzzy logika olyan logikaforma, amely közelítő, nem pedig rögzített és pontos következtetésekkel foglalkozik. Ez a Boole-logika kiterjesztése, amely a részleges igazság fogalmát kezeli – az 'teljesen igaz' és 'teljesen hamis' közötti igazságértékeket." [3]



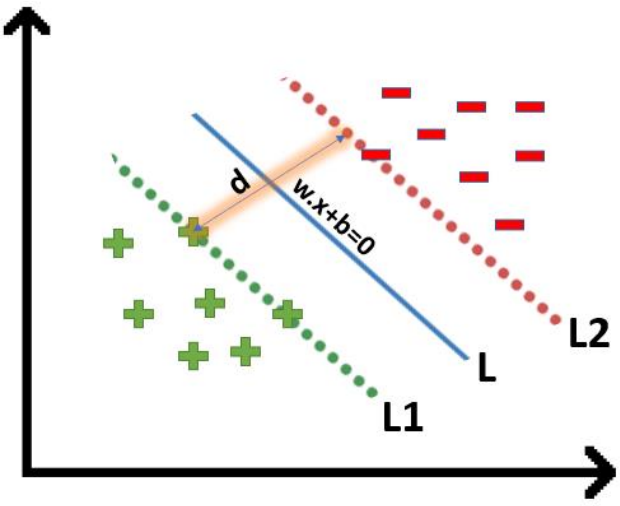
3.4 Fuzzy Logika alapú kockázatbecslés

A fuzzy alapú kockázatelemzés olyan módszer, amely a fuzzy logikát alkalmazza a kockázatok felmérésére és kezelésére, különösen bizonytalanságok és pontatlan adatok esetén. Ez a megközelítés a fuzzy halmazelméletet használja a kockázatok kvantifikálására és elemzésére a homályosság és kétértelműség kezelése révén, lehetővé téve a rugalmasabb és átfogóbb kockázatértékelést a hagyományos módszerekhez képest. [4]

Például a fuzzy kockázatelemzés egy esemény kockázatát nem egyszerűen "magasnak" vagy "alacsonynak" értékeli, hanem ezekbe a kategóriákba való tagság mértékeként, lehetővé téve a finomabb döntéshozatalt. [4]

3.5 Support Vector Machine

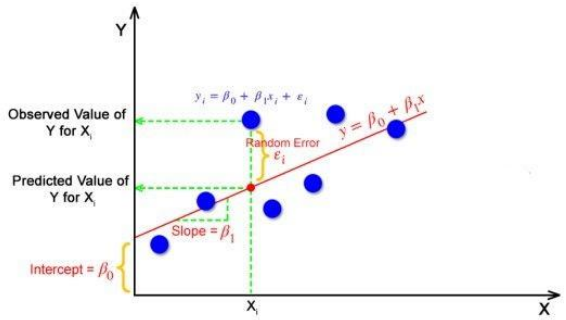
A támogatott vektorgép (Support Vector Machine, SVM) egy felügyelt tanulási algoritmus, amelyet elsősorban osztályozási és regressziós problémák megoldására használnak. Az SVM célja egy optimális választóvonal (hipersík) megtalálása, amely maximális távolságot (margin) tart fenn a különböző osztályokhoz tartozó adatok között. Az optimális hipersík az, amely a legjobban elválasztja az osztályokat, és a legközelebbi adatok (támogató vektorok) minimális távolságát maximalizálja a hipersíktól. [13]



3.6 Lineáris Regressió

A lineáris regresszió az egyik legismertebb és leggyakrabban használt statisztikai módszer az adatelemzésben. Célja, hogy két vagy több változó közötti kapcsolatot modellezze és kvantifikálja. Az alapvető cél

az, hogy egy egyenes vonallal (lineáris modellel) írja le az adatok közötti kapcsolatot. A lineáris regresszió matematikailag az alábbi egyenlettel írható le: [6][7][8]



3.7 *Neurális hálók*

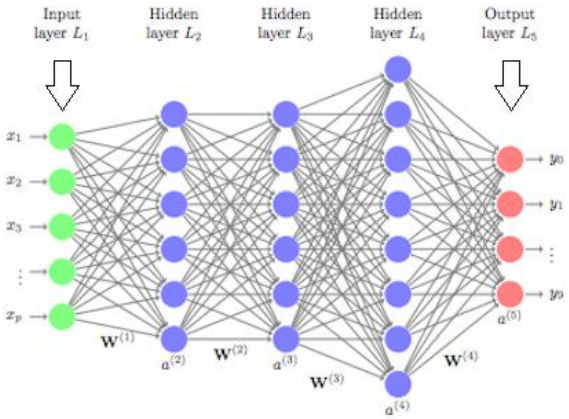
A neurális hálók a gépi tanulás egy speciális formája, amely az emberi agy működését mintázza. Ezek a hálók az adatfeldolgozási és mintafelismerési feladatokhoz kiválóan alkalmazhatók. A neurális hálók alapegysége a neuron, amelyet mesterségesen modelleznek, hogy a bemeneti jelekre adott kimeneteket képezzen. A neurális háló több rétegből állhat, ezek közül a legfontosabbak a bemeneti réteg, a rejtett rétegek és a kimeneti réteg. [9][10]

A neurális hálók matematikai modellje alapvetően az alábbiak szerint írható le:

Bemeneti réteg: Az adatok ebbe a rétegbe lépnek be.

Rejtett rétegek: Ezek a rétegek az adatok bonyolultabb feldolgozását végzik, itt történik az adatokból történő minták felismerése.

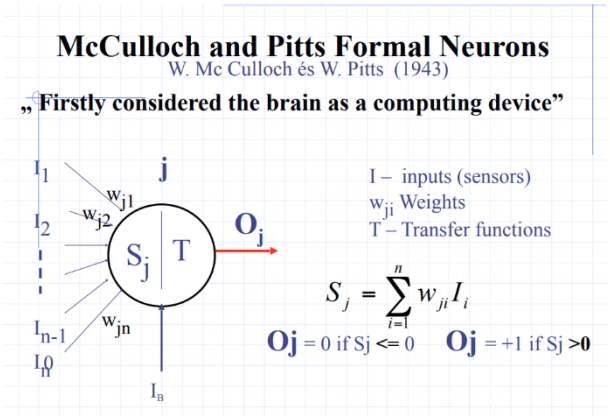
Kimeneti réteg: Az utolsó réteg adja meg a végső előrejelzést vagy osztályozást. [9][10]



1. ábra Neuron háló ábrázolása [11]

3.8 *Neuronok*

McCulloch és Pitts formális neuronjának modellje a mesterséges intelligencia alapvető építőköve, amely a biológiai neuronok működését imitálja. A modell bemeneti jelekből áll, amelyek mindegyike egy súllyal van összekötve a neuronhoz, és a bemeneti jelek súlyozott összegét egy transzferfüggvény (gyakran lépcsőfüggvény) segítségével alakítja át kimeneti jellé. A kimenet akkor aktiválódik, ha a súlyozott összeg meghalad egy előre meghatározott küszöbértéket, ami az alapvető logikai műveletek megvalósítását teszi lehetővé.



2. ábra Neuron felépítése [12]

4 ADATBÁZIS BEMUTATÁSA

Az EuRepoC 2000-es évekig visszamenőleg napjainkig nyújt adatok, amik nyílt forrásból elérhetőek.[5]

4.1 *Az adatfájlok szerkezete*

A letölthető Excel fájl három lapra tagolódik, a CSV letöltés pedig három külön fájlt tartalmaz. [5]

4.2 *Fő adatkészlet*

Ez a fájl vagy lap tartalmazza az egyes incidensekhez kódolt összes változót, úgy rendezve, hogy egy sor egy eseménynek feleljen meg – a mi fő vizsgálati egységünknek. Ha egyetlen eseményhez egyetlen változóhoz több kód is tartozik, ezeket pontosvesszővel választják el ugyanazon a cellán belül. [5]

4.3 *Vevő adatkészlet*

Ebben a fájlban vagy lapon az érintett entitások és személyek (fogadók) adatai átstrukturálva vannak az elemzés megkönnyítése érdekében. Minden cella csak egyetlen kódot tartalmaz, az adatok több sorban vannak „kicsomagolva”. Így egyetlen incidens több sort is átívelhet, amelyek az egyes eseményekhez rendelt egyedi azonosítók révén azonosíthatók. [5]

4.4 *Hozzárendelési adatkészlet*

Ez a lap vagy fájl a vevő adatkészletéhez hasonló megközelítést követ. A hozzárendelési adatok több sorban „kicsomagolva” vannak, így minden cella csak egy kódot tartalmazhat. Itt is egyetlen incidens több sort is elfoglalhat, és az egyedi azonosító lehetővé teszi az egyes események egyszerű nyomon követését. [5]

4.5 *Gyűjtött adatok*

Az EuRepoC az alábbi adatokat gyűjtötte:

1) ID
2) name
3) description
4) start_date
5) end_date
6) inclusion_criteria
7) inclusion_criteria_subcode
8) source_incident_detection_disclosure

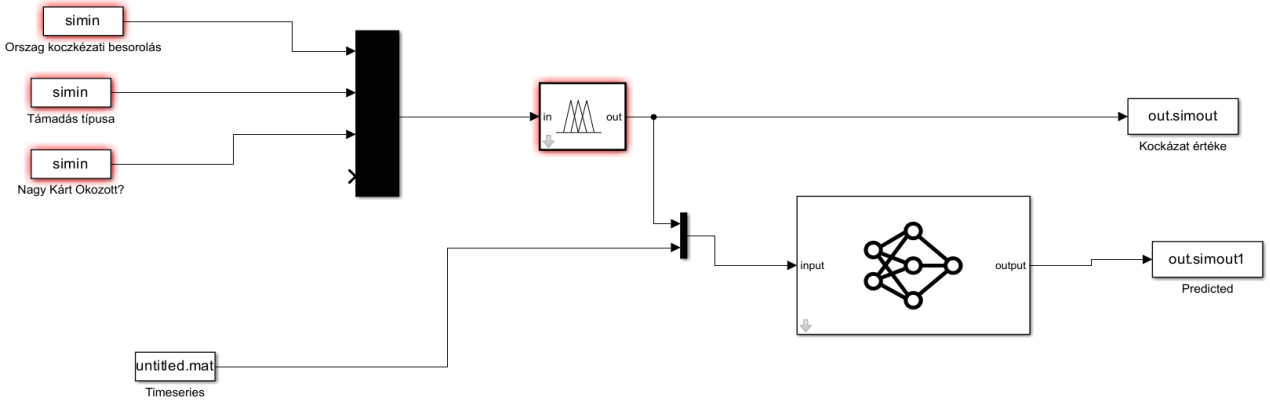
9) incident_type
10) receiver_name
11) receiver_country
12) receiver_region
13) receiver_category
14) receiver_category_subcode
15) initiator_name
16) initiator_country
17) initiator_category
18) initiator_category_subcode
19) number_of_attributions
20) attribution_ID
21) attribution_date
22) attribution_type
23) attribution_basis
24) attributing_actor
25) attribution_it_company attributing_country
26) attributed_initiator
27) attributed_initiator_country
28) attributed_initiator_category
29) sources_attribution cyber_conflict_issue
30) offline_conflict_issue
31) offline_conflict_issue_subcode
32) offline_conflict_intensity
33) offline_conflict_intensity_subcode
34) number_of_political_responses
35) political_response_date
36) political_response_type
37) political_response_type_subcode
38) political_response_country
39) political_response_actor zero_days
40) zero_days_subcode
41) MITRE_initial_access
42) MITRE_impact
43) user_interaction
44) has_disruption
45) data_theft
46) disruption
47) hijacking
48) physical_effects_spatial
49) physical_effects_temporal
50) unweighted_cyber_intensity
51) target_multiplier
52) weighted_cyber_intensity
53) impact_indicator
54) impact_indicator_value
55) functional_impact
56) intelligence_impact
57) political_impact_affected_entities
58) political_impact_affected_entities_exact_value
59) political_impact_third_countries
60) political_impact_third_countries_exact_value
61) economic_impact
62) economic_impact_exact_value
63) economic_impact_currency
64) state_responsibility_indicator
65) IL_breach_indicator
66) IL_breach_indicator_subcode
67) evidence_for_sanctions_indicator
68) number_of_legal_responses
69) legal_response_date
70) legal_response_type
71) legal_response_type_subcode
72) legal_response_country legal_response_actor
73) legal_attribution_reference
74) legal_attribution_reference_subcode
75) legal_response_indicator
76) casualties
77) sources_url
78) added_to_DB
79) updated_at

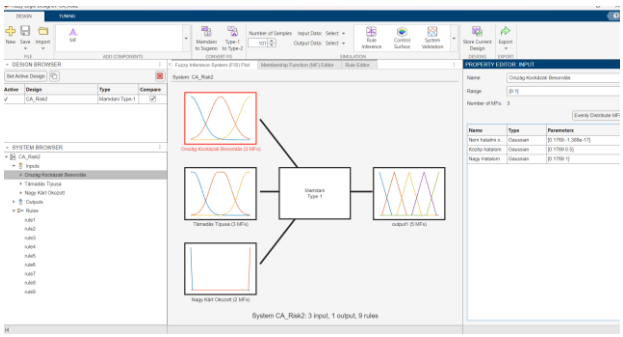
5 FUZZY LOGIKA ALAPÚ KOCKÁZAT BECSLŐ RENDSZER

Készítettünk egy Fuzzy logika alapú kockázat becslő rendszert a Matlab Fuzzy Logic Designer Toolbox használatával. A rendszerünk 3 értéket vesz figyelembe az adathalmazból. Első az ország kockázati besorolása. Ez azért fontos tényező, mivel egy nagyobb kibertámadás esetén a nagy és középhatalmaknak lehetősége van nem csak a kibertérben, hanem például pénzügyi, jogi vagy diplomáciai szankciókkal sújtani a vélt támadót (azért csak véltet, mivel a támadások vissza nyomozása sokszor technikai okok miatt lehetetlen). 3 kategóriát alkalmaztunk, az egyszerűség kedvéért. Nagyhatalmak, középhatalmak és kisebb országok. Nagyhatalmaknak az alábbiakat csoportosítottuk: USA, Kína, Oroszország, Egyesült Királyság, Franciaország, Németország. Középhatalmak esetében: India, Japán, Brazília, Kanada, Ausztrália, Olaszország, Dél-Korea, Törökország, Spanyolország, Mexikó. Minden más esetben kisebb ország csoportot alkalmaztuk. A csoportosítás során figyelembe vettük az országok gazdasági teljesítményét, katonai erőit, politikai befolyását és a kulturális befolyásosságuk. A tagsági függvényeket Gaussi típusúra állítottuk és egyenlően osztottuk el őket.

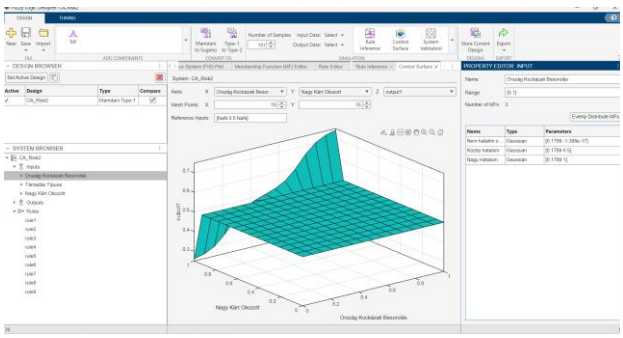
A második bemeneti érték a támadás típusa. Itt is három részre osztottuk fel a bemenetet: kis kiterjedésű támadás, közepes kiterjedésű támadás, nagy kiterjedésű támadás. Tagsági függvénynek két oldalú Gaussi függvényeket használtunk egyenletesen elosztva a tartományon.

Az utolsó bemeneti értékünk egy bool változó ahol annyi információt hordoz a bemenet, hogy maga a támadás kiterjedtségtől függetlenül nagy kárt okozott-e? Az EuRepoC szerint érhetik a kormányzati oldalakat, kritikus infrastruktúrákat, médiát, politikai pártokat és még másokat is a támadások. Ezek mind kihatással lehetnek az ország állampolgárainak életére.

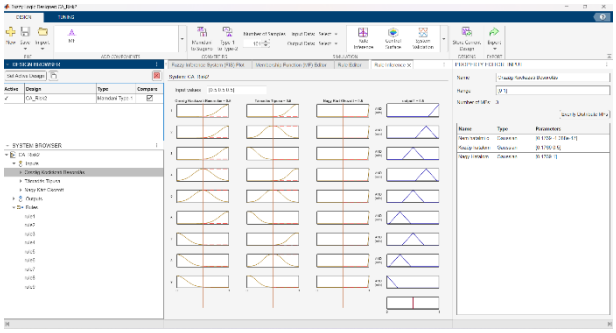




3. ábra Kép a Fuzzy Interfaceről | Szerző által szerkesztett



4. ábra Kép a Fuzzy Interfaceről Controll Surface| Szerző által szerkesztett



5. ábra Kép a Fuzzy Interfaceről Rule Inference| Szerző által szerkesztett

6 PREDIKCIÓS RENDSZER

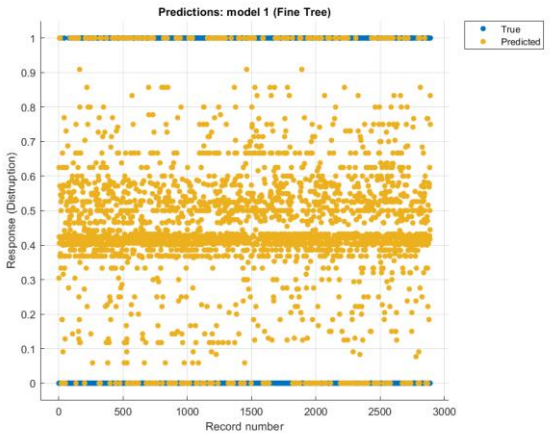
6. ábra Predikciós rendszer | Szerző által szerkesztett

6.1 Modell és Mérések bemutatása

Predikciók készítéséhez a Matlab Regression Learner Toolboxát használtuk. A mérések közül az alábbi 3-at szeretném ismertetni.

6.1.1 Fine Tree

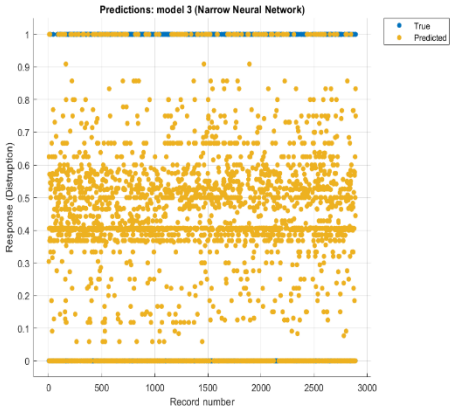
A Fine Tree algoritmust 4-es minimum levél mérettel futattuk le. Az RMSE validáció: 0,49035



7. ábra Fine Tree Response Plot| Szerző által szerkesztett

6.1.2 Narrow Neural Network

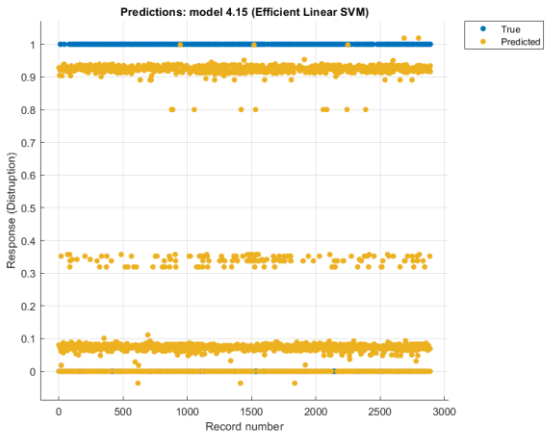
A Narrow Neural Networkot 1 teljesen csatlakoztatott réteggel. 10.-es bemeneti réteggel ReLU aktiváció függvénnyel és 1000-es iterációs limittel futtatok le. Az RMSE validáció értéke 0,49623.



8. ábra Narrow Neural Network Plot| Szerző által szerkesztett

6.1.3 Efficent Linear SVM

Az Efficient Linear SVM-et automata üzemmódban futattuk le. RTC értéke: 0,0001. RMSA validáció: 0,5787.



9. ábra Efficient Linear SVM Response Plot | Szerző által szerkesztett

IRODALOMJEGYZÉK

[1] https://eurepoc.eu/about-us/ Szerző által fordított

[1] Murphy, Kevin P.. "Machine learning - a probabilistic perspective." Adaptive computation and machine learning series (2012).

[2] J. M. Mendel, "Fuzzy logic systems for engineering: a tutorial," in Proceedings of the IEEE, vol. 83, no. 3, pp. 345-377, March 1995, doi: 10.1109/5.364485.

[3] Omidvar, M., Zarei, E., Ramavandi, B., Yazdi, M. (2022). Fuzzy Bow-Tie Analysis: Concepts, Review, and Application. In: Yazdi, M. (eds) Linguistic Methods Under Fuzzy Information in System Safety and Reliability Analysis. Studies in Fuzziness and Soft Computing, vol 414. Springer, Cham. https://doi.org/10.1007/978-3-030-93352-4_3

[4] https://eurepoc.eu/database/

[5] *European Repository of Cyber Incidents (EuRepoC) (2024) "Global Dataset of Cyber Incidents V.1.2". doi: 10.5281/zenodo.11108195.*

[6] Montgomery, D.C., Peck, E.A., & Vining, G.G. (2012). Introduction to Linear Regression Analysis. John Wiley & Sons.

[7] Weisberg, S. (2005). Applied Linear Regression. John Wiley & Sons.

[8] Kutner, M.H., Nachtsheim, C.J., & Neter, J. (2004). Applied Linear Regression Models. McGraw-Hill Irwin.

[9] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

[10] Bishop, C.M. (2006). Pattern Recognition and Machine Learning. Springer.

[11] https://www.researchgate.net/figure/A-sample-neural-network-layout-with-three-hidden-layers_fig5_368540750

[12] Dr. Kutor László előadása OE NIK 2023 https://elearning.uni-obuda.hu/main/pluginfile.php/1132079/mod_resource/content/0/IS%202023-2-2.pdf

[13] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297.