

Adatelemzés R környezetben - Edutus Egyetem Passzív Ház

1. BEVEZETÉS	1
2. A MESTERSÉGES INTELLIGENCIA, A GÉPI TANULÁS ÉS A MÉLYTANULÁS	1
3. A KORRELÁCIÓ ÉS REGRESSZIÓ	6
4. AZ R PROGRAMNYELV	9
5. AZ ADATOK ELŐKÉSZÍTÉSE ÉS ELEMZÉSE	10
6. ÖSSZEFOGLALÁS	14
IRODALOMJEGYZÉK	15
MELLÉKLETEK.....	17

Ábrák és táblázatok jegyzéke

Ábra 1 A mesterséges intelligencia, gépi tanulás, mélytanulás kapcsolata.....	3
Ábra 2 Mesterséges neurális hálózat felépítése.....	5
Ábra 3 Microsoft Azure Cheat Sheet	6
Ábra 4 Diagram mátrix	17
Ábra 5 Egyváltozós lineáris korreláció modell	18
Ábra 6 Kétváltozós lineáris modell	18
Ábra 7 Kétváltozós lineáris modell kereszthatással	19
Ábra 8 Neurális hálózat tíz rejtett réteggel 2 prediktor változóval	19
Ábra 9 Neurális hálózat tíz rejtett réteggel 12 prediktor változóval	20
Ábra 10 A modellek összehasonlítása	20
Ábra 11 R szkript a 0-24 óra alatt rögzített adatok beolvasásához	27
Ábra 12 R szkript a 9:45 - 10:00 között rögzített adatok beolvasásához.....	34
Ábra 13 A különböző modellek összehasonlítása.....	54
Táblázat 1 A vizsgált modellek és változóik.....	12
Táblázat 2 A változók rövidítéseinek a jelentése	12
Táblázat 3 Az előre jelzett napi értékek és a valóságban rögzített értékek eltérései.....	13

1. Bevezetés

A megújuló energiatermelés a megújuló energiák hasznosítása során olyan problémák kerülnek előtérbe, mint például a nehezen kontrollálhatóság és előre tervezhetőség. Egy megfelelő teljesítmény előrejelzés sokat segítene például a karbantartások ütemezésében, illetve az energia mennyiségek beérkező energia mennyiségek kontrollálásában, energiaforrások hasznosítási módjainak a versenyképességén. Manapság a mesterséges intelligencia szinte az élet minden területén megtalálható, az orvosi műszerek, tőzsdei előrejelzések, vállalati irányítási és tervező programok, gépi fordítás, kiberbiztonsági termékek fejlesztése és az okostelefonunkban az öntanuló algoritmusok és sorolhatnám a példákat tovább. Ezt a tervezhetőséget vizsgálatuk meg az Edutus Egyetem projektfeladat tantárgyán belül, ahol megismerkedtünk az R programnyelv alapjaival, illetve megnéztük alkalmazható e a napi hőmennyiség előre jelzésére az Edutus Passzívház Mariann rendszerének gyűjtött adatai alapján.

2. A mesterséges intelligencia, a gépi tanulás és a mélytanulás

Mi az a mesterséges intelligencia? Ezt a kifejezést olyan gépekre, robotokra és szoftverekre alkalmazzuk, melyek komplex feladatokat előre beprogramozás nélkül önállóan el tudnak végezni. Ezen számítógép-architektúra képes arra, hogy olyan emberre jellemző kognitív funkciókat tud lemásolni, mint amilyen a tanulás és a problémák megoldása. (Microsoft, 2022) A kutatók és szakemberek szerint a mesterséges intelligencia még a közeljövőben nagyon meg fogja változtatni az életünket. (Pumhösel, 2019)

Jelenleg a fejlődés nagyon gyorsan történik, de már viszonylag régebben elkezdődött. A mesterséges intelligencia történetének kezdete az 1950-es évekre tehető, ebben az időben Alan Turing brit informatikus kidolgozott egy tesztet annak bizonyítására, hogy egy számítógép képes-e úgy gondolkodni, mint egy ember. A teszt során bírálók tesznek fel kérdéseket két tesztalanyra, ahol a két alany közül az egyik gép a másik pedig ember. Az, hogy a mesterséges rendszer megválaszolta a kérdéseket, az nem jelenti azt, hogy meg is értette azt. A megértés tényét szűrik ki a teszttel.

A kutatók az ötvenes az években alkották meg a mesterséges intelligencia kifejezést is. Ezután intelligens humanoid robotok jelentek meg könyvekben és filmekben. Ebben az időben még

úgy gondolták, hogy ezeknek a gépeknek nagynak és minden általános problémát megoldónak kell lennie (Think big). A hatvanas években a Massachusettsi Műszaki Egyetemen először oldotta meg egy számítógép egy IQ-teszt feladatait. Röviddel ezután elkészül az első olyan program, amely önállóan párbeszédet képes volt folytatni emberekkel.

Az első önállóan navigáló robot neve Shakey volt. 1966 és 1972 között hozták létre a kaliforniai Stanford Kutatóintézetben, ez a robot az akkori kor legmodernebb érzékelőivel volt felszerelve és a vezérlési feladatokat már mesterséges intelligencia segítségével oldotta meg. A robotot fő feladata a problémamentes közlekedés volt.

Az következő az ALVINN mesterséges intelligencia rendszer, mely 1989-ben egy önvezető autót irányít az USA keleti partjáról a nyugati partra. Ez rendszer már egy három szintű neurális hálózatot használt. Az autóban lévő chipnek nem volt nagyobb teljesítménye, mint egy Apple okosórának. 1989-hez kötődik szintén, hogy az IBM ekkor alkotta meg a Deep Blue szuperszámítógépet, a gépnek sokéves tanulásra volt szüksége ahhoz, hogy legyőzze az akkori sakkvilágbajnokot, Kaszparovot. A gép a paramétereket saját maga határozta meg, rengeteg mesterverseny lépéseinek elemzésével.

A 1997 – ben indult el a Google keresőmotorja és hamarosan monopolhelyzetbe került a többi internetes keresővel szemben, köszönhetően annak, hogy mesterséges intelligenciát alkalmaztak.

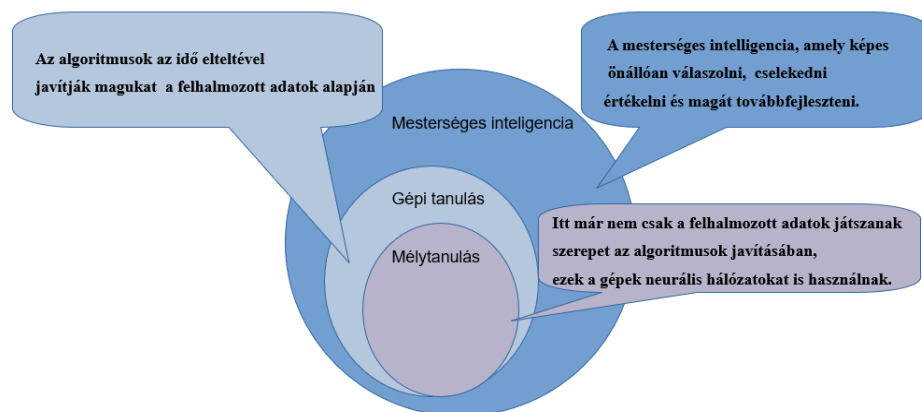
A 2010-es évektől kezdve a fejlődés felgyorsult. A Google önvezető autója, melynél már nem nagyon tudunk különbséget tenni, hogy ember, vagy gép vezeti e, 2014-ben került bemutatásra, majd a szintén a Google fejlesztése az Alpha Go megveri az a világ legjobb Go játékosát. A Go-t összetettsége miatt nem lehet algoritmusokkal leképezni, melyek minden lehetséges lépést összevetnek és súlyoznak.

Napjainkra eljutottunk odáig, hogy mesterséges intelligencia segítségével hoznak létre egy másik mesterséges intelligenciát, illetve a mesterséges intelligencia a mindennapi életünk része lett (Vogels, 2018).

Az emberek naponta használnak olyan asszisztenseket, mint az Apple Siri, az Amazon Alexa, vagy a Google Now. A gépi tanulás mindegyikben ilyen alkalmazásban főszerepet játszik. Az online fordítók, a képfelismerő szoftverek és hamarosan az önvezető autók is a napi életünk részévé fognak válni.

De mi is az a gépi tanulás? A kifejezés első definíciója 1959-ből Artur Samueltól származik, aki a gépi tanulást olyan tudományterületként írta le, ahol a számítógépek képessé teszik tanulásra, anélkül, hogy erre kifejezetten programoznák őket (Bucioglu, 2021).

A nagy mennyiségű adatok elemzése statisztikai módszerekkel és algoritmusokkal teszik lehetővé azt, hogy a gépek képesek legyenek önállóan „tanulni”, a megtanult mintákat felhasználva adatalapú, intelligens döntéseket hozni. A programozott funkciókkal szemben a gépi tanulás nagy előnye az adaptivitás, tehát az új szituációkhoz való alkalmazkodás és az ezzel járó komplexitás. Ezért gépi tanulás a mesterséges intelligencia egy részhalmazának tekinthető (Wuttke, 2022).



Ábra 1 A mesterséges intelligencia, gépi tanulás, mélytanulás kapcsolata

Forrás: Saját ábra

A gépi tanulás (Machine Learning, vagy csak egyszerűen ML) tulajdonképpen egy gyűjtőfogalom a tapasztalatból, adatokból származó tudás mesterséges kifejezésére. Ahhoz, hogy működhessen, a mesterséges intelligencia szoftverének szüksége van az adatokra, algoritmusokra és különböző előre megadott szabályokra, majd ezeket önállóan értelmezve mintákat és összefüggéseket kell, hogy felismerjen. Az adatok alapján bizonyos események bekövetkezési valószínűségei előre meghatározhatók és ezen valószínűségek súlyozása alapján már döntéseket lehet hozni.

Ez a rendszer csak a mintapéldák mintázata alapján tanul és a tanulási fázis befejezése után általánosít, majd felépít egy statisztikai modellt. Ami a gépi tanulás szempontjából fontos, az a nagy mennyiségű adat megléte és hozzáférhetősége. Minél nagyobb mennyiségű olyan

mintaadat áll rendelkezésre, mely a probléma megoldásához relevánsan kötődik, akkor annál pontosabb modell, mely a mintázatokra felépíthető.

Ez az úgy nevezett „Big Data” fogalma, mely kifejezést nem szoktuk magyar nyelvre lefordítva használni, de ha megtesszük akkor sok, vagy rengeteg adatot jelent. A valós jelentése ennél jóval összetettebb, mert a nagy mennyiségű adathoz hozzá is kell férni, fel kell dolgozni, lehetőleg a leggyorsabb módszerrel úgy, hogy ezek az adatok közben folyamatosan változnak, gyarapodnak.

A „Big Data”, mint fogalom csak a 2000-es évek elején jelent meg a köztudatban, amikor Doug Laney iparági érik el egymást és kommunikálnak egymással) elterjedése óta az adatfolyamok soha nem látott sebességgel érkeznek az adatbázisokba, és ezeket azonos időben (real time) kell feldolgozni.

Változatosság (Variety) Az adatok a legkülönbözőbb formátumokban kerülhetnek adatbázisba, lehetnek e-mailek, videók, képek, hangok, szöveges dokumentumok, numerikus adatok, vagy pénzügyi és tranzakciós elemző a 3-V modelljében megfogalmazta a Big Data ma is elfogadott definícióját:

Mennyiség (Volume): az adatok gyűjtése számos forrásból történhet, üzleti tranzakciókból, intelligens eszközökből, ipari berendezések és gépekből, képekről, videókról, közösségi médiából. Ezeknek az adatoknak tárolása a múltban eléggé problémás volt, de már leegyszerűsödött a különböző költséghatékony tárolási platformoknak köszönhetően.

Sebesség (Velocity): az adatgenerálás és -feldolgozás sebessége, az adatok többsége különböző eszközök internet alapú kapcsolatán kerül továbbításra (IoT, ahol eszközök adatokat továbbítanak egymás között, a dolgok internete) (COSO IT, 2017).

Sokan használnak még további két „V”-t a definíció meghatározásakor:

Veracity, a valódiság, az adatok fel használhatóságának bizonytalansága és tulajdonképpen a minősége. Az adatok nem használhatóak fel közvetlenül, hanem újra fel kell dolgozni őket. Ez extra költséggel jár a vállalatok számára.

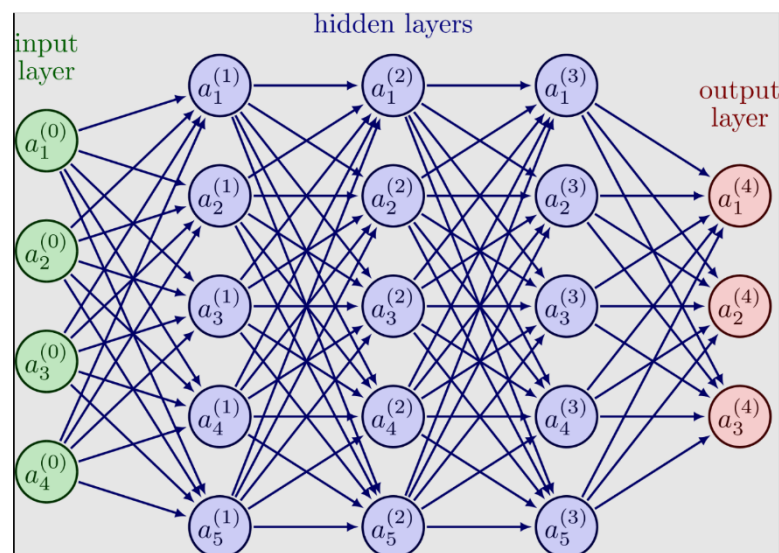
A **Value,** a nagy (összekapcsolt) adatmennyiségek által generált hozzáadott érték, vagy üzleti érték, és természetesen a gépi tanulási technikák alkalmazhatósága ezekre a hatalmas adatbázisokra. Ez valószínűleg az egyik legfontosabb érv a Big data alkalmazásának szempontjából, mert érték nélkül nem lenne értelme ezekkel a projektekkel foglalkozni (Anil Jain, 2016).

A Big Data alkalmazása erőteljesen növekvő trendet mutat napjainkban, hiszen a fentiekből láthatjuk, hogy erőteljes értéknövelő hatása van a vállalatra nézve, már csak visszautalva a klasszikus idézetre: az „információ hatalom, ha jól használod győzelem”.

A mélytanulás a gépi tanulás egy speciális részterülete. Nagy mennyiségű adat feldolgozásával és az algoritmusok önálló javításán kívül utánozza az emberi tanulási viselkedést. Egy "mélytanulási algoritmusban", vagy "mély neurális hálózatban" számos köztes, úgynevezett rejtett réteg lehet a bemeneti és kimeneti adatok között.

A gép mélytanulás során a neurális hálózatokat használja az algoritmusaihoz, a minták megfelelő felismeréshez. A neurális hálózati struktúrák napjainkban csak felületesen hasonlítanak az emberi agyhoz, de a mélytanulás sikerességének titka az emberi agyhoz való hasonlóság, hogy a különböző rétegekből álló hálózatokat képez, hoz létre és ezeket tanítják be különböző mintázatokra. Ezek a bemenet és kimenet közötti bonyolult struktúrák teszik lehetővé a minták felismerését és az új jelenségekre vonatkozó előrejelzéseket és predikciókat.

Ez a mélytanulási képesség az emberi gondolkodásra jellemző mindenre kiterjedő ismeretszerzéssel ellentétben csak egy területre korlátozódik. Tehát a rendelkezésre álló adatoknak, melyeket vizsgálni szeretnénk szorosan kapcsolódniuk kell ahhoz a problémához melyet meg szeretnénk oldani (Wuttke, 2022).



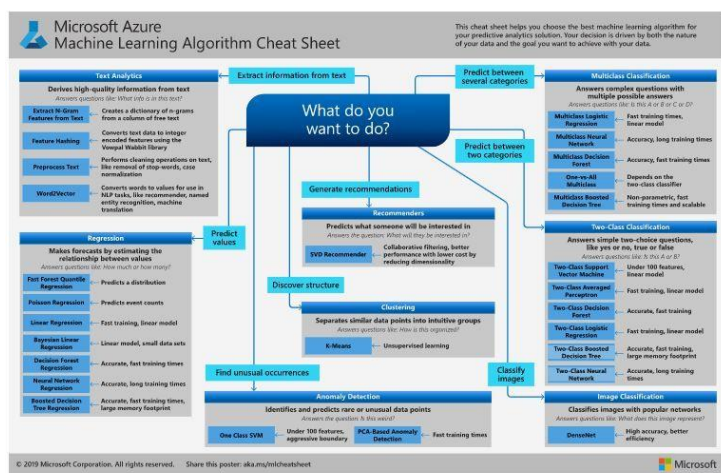
Ábra 2 Mesterséges neurális hálózat felépítése

Forrás: (Neutelings, 2021)

Input layer: bemeneti réteg, Hidden layer: rejtett réteg, Output layer: kimeneti réteg

A gépi tanulás nagyon sokféle területen alkalmazható, ahol rendelkezésre állnak megfelelő adatok. Egy konkrét feladat megoldása előtt el kell döntenünk melyik típusú gépi tanulást választjuk. Ebben a dolgozatban adatokat előre szeretnénk jelezni, ezért felügyelt gépi tanulást fogom alkalmazni, mivel az adatokat először meg kell tisztítani és a jobb konfiguráció érdekében ki kell választani az algoritmusok paramétereit. Feltételezésünk is azon alapul, hogy egy-, vagy többváltozós lineáris regresszió alkalmazásával összefüggést találunk az adatok között és ezáltal a lehetséges kimeneti eredmények valószínűségét meg tudjuk becsülni.

A Microsoft Azure Cheat Sheet-je segítséget ad a megfelelő módszer kiválasztásában.



Ábra 3 Microsoft Azure Cheat Sheet

Forrás: (Microsoft, 2021)

3. A korreláció és regresszió

Az eredeti feltevésemre keresve a választ, a változók lehetséges összefüggésének vizsgálatát statisztikai módszerekkel és az R- környezet segítségével fogom végrehajtani. A statisztikai módszerek a korreláció és regresszióanalízis.

A korrelációanalízis során két, vagy több változó kapcsolatát vizsgáljuk, tehát a korreláció megmutatja, hogy a változók között van-e kölcsönös, oda visszaható kapcsolat. A korrelációs számítás során az X és Y véletlen változók közötti kapcsolat erősségét és irányát. Egyszerűen fogalmazva, a korrelációelemzés kiszámítja, hogy az egyik változó milyen mértékben változik a másik változó változása hatására. A magas korreláció a változók közötti erős kapcsolatot jelzi, míg az alacsony korreláció azt jelenti, hogy a változók gyengén függenek

egymástól. A korrelációelemzés tehát két változó, vagy adathalmaz közötti kapcsolatok, minták, jelentős összefüggések és tendenciák azonosítására használható (dr.Némethné Gál, 2000).

A korrelációs együttható az a számérték, amely a korreláció típusát, azaz a két változó közötti statisztikai kapcsolatot jelzi. A korrelációs együttható (r) értéke + 1 és - 1 között változik a változók közötti kapcsolat erősségét tekintve. Minél közelebb van a korrelációs együttható értéke a 0-hoz, annál gyengébb a kapcsolat a két változó között. A kapcsolat irányát a korrelációs együttható előjele jelzi; a „+” jel közvetlen kapcsolatot, a „-” jel pedig fordított kapcsolatot jelez (Molnár, 2021).

Pozitív korreláció: A két változó közötti pozitív korreláció azt jelenti, hogy mindkét változó ugyanabba az irányba mozog. Az egyik változó növekedése a másik változó növekedéséhez vezet, és fordítva. Ha például több időt töltünk futással, több kalóriát égetünk el.

Negatív korreláció: A két változó közötti negatív korreláció azt jelenti, hogy a változók ellentétes irányban mozognak. Az egyik változó növekedése a másik változó csökkenéséhez vezet, és fordítva. Például az autó sebességének növelése csökkenti a célállomás eléréséhez szükséges időt.

Gyenge/nulla korreláció: Nincs korreláció, ha az egyik változó nincs hatással a másikra. Például nincs összefüggés aközött, hogy egy személy hány évig járt iskolába, és hogy hány betűből áll a vezetékneve.

Az ismérvek kapcsolatának erősségét a Pearson féle korrelációs együttható fejezi ki- jele: r.

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Ha r megközelítőleg nulla, akkor a változók nem korrelálnak egymással.

Vagyis nincs lineáris kapcsolat közöttük. Ha az r értéke nagyobb, mint 0, akkor pozitív korrelációról beszélünk. Ez azt jelenti, hogy az x változó növekedése az y növekedésével jár. Végül megvizsgáljuk azt az esetet, amikor az r értéke nulla alatti, azaz negatív korrelációról van szó. A pozitív korrelációval ellentétben itt az y értékei csökkennek, ha az x értékei nőnek (Molnár, 2021).

A regresszioelemzés egy statisztikai elemzési módszer. A regresszió segítségével megvizsgálható, hogy egy, vagy több másik változó értékeivel mennyire jól lehet megjósolni egy változó értékeit. Ehhez meg kell vizsgálni a változók korrelációját, és ez alapján létrehozni egy trend függvényt. Minél erősebb a kapcsolat a változók között, annál jobban meg lehet jósolni az egyiket a másik alapján.

A megjóslandó változót függő változónak nevezzük. A függő változó előrejelzésére használt változót viszont független változónak, vagy prediktornak nevezzük. A regresszioelemzést különböző célokra használják. Az új értékek előrejelzésén túl a különböző változók közötti kapcsolatok részletesebb vizsgálatára is szolgál. A módszer során az adatpontjainkra legjobban illeszkedő egyenest keressük.

A lineáris regresszió a kétváltozós regresszió legegyszerűbb alakja. A lineáris regresszió során egy, vagy több másik változó segítségével próbáljuk megjósolni egy változó értékét.

A legegyszerűbb eset a síkon pontokra illeszkedő egyenes, az egyváltozós lineáris regresszió:

$$y = \beta_0 + \beta_1 * x$$

A következő szintű esetben a pontokra már egy sík illeszkedik, a többváltozós lineáris regresszió:

$$y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2$$

A többváltozós regressziós modell egyenlete:

$$y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \dots + \beta_p * x_p$$

(Molnár, 2021)

A nem lineáris regressziónál nem áll fent lineáris függvénykapcsolat a változók között. A nem lineáris regresszió alaptípusai:

Logaritmikus:

$$y = a + b * \ln(x)$$

Exponenciális:

$$y = a * b^x$$

Hiperbolikus:

$$y = a + b/x$$

Hatványkitevős:

$$y = a * x^b$$

Polinominális:

$$y = a + (b_1 * x) + (b_2 * x_2) + \dots + (b_n * x_n)$$

(SPSSABC.HU, 2021)

4. Az R programnyelv

Az R egy objektumorientált, interaktív programozási nyelv, amellyel statisztikai kiértékelések végezhetők. A statisztikai elemzések könnyen elvégezhetők, változatos grafikák készíthetők és szimulációk végezhetők. Az R nyílt forráskódú (azaz a forráskód szabadon hozzáférhető és bárki által használható), ingyenes és különböző operációs rendszerekhez elérhető. Ezenkívül az R-t rendszeresen frissítik (az "R Core Team" által). Így mindig naprakész és hibamentes marad, és folyamatosan optimalizálják. Az 1970-es évek közepén az S programozási nyelvet az Egyesült Államokban fejlesztette ki a John Chambers és munkatársai (Bell Laboratories) statisztikai vizsgálatok elvégzéséhez, illetve szimulációk és grafikák készítéséhez (TIBCO Software Inc., 2022). Az R egy interpretált szkript nyelv, a parancsok nem kerülnek lefordításra, hanem az R – értelmező egyesével értelmezi és soronként végrehajtja azokat. (Molnár, 2021) Nagyon sokféle művelet hajtható végre az R segítségével, kezdve a legegyszerűbb műveletektől a nagyon bonyolult statisztikai számításokig.

Az R nyelv segítségével az adatok nagyon rugalmasan, gyorsan értékelhetőek ki és jeleníthetőek meg. Ezért az alkalmazás ideális az olyan feladatokhoz, mint a prediktív analitika, vagy az adatbányászat. Nagyvállalatok is, mint például az Oracle vagy a Microsoft is használják ezt a programozási nyelvet. Az R könnyen integrálható. Számos interfész és integrációs lehetőség elérhető a piacon (Litzel & Dipl.-Ing. (FH) Luber, 2018). Az RStudio az R statisztikai szoftver integrált fejlesztőkörnyezete. Egy IDE különböző eszközöket kínál, amelyek megkönnyítik a munkát szoftverekkel. Az analízist RStudio-ban fogom végrehajtani.

5. Az adatok előkészítése és elemzése

Az Edutus passzívház napkollektor rendszerébe beépített két besugárzásmérő (Q1, Q2), a hőmérő, amely a külső hőmérsékletet méri, és a szél erősség mérő adatai másodpercenkénti mintavételezési idővel rögzítésre kerülnek a MARIANN program segítségével. Megvizsgáljuk, hogy a rendelkezésre álló adatok reggel 9:45 és 10:00 között beérkezett adatok alapján lehetséges-e a napi várható besugárzás (időjárás) előrejelzése.

Tehát a felállított hipotézisek a következők:

Nullhipotézis:

H_0 : Reggel 9:45 és 10:00 között beérkezett adatokból nem jósolható meg, hogy mekkora lesz a napi hőtermelés

Ellenhipotézis:

H_1 : Reggel 9:45 és 10:00 között beérkezett adatokból megjósolható, hogy mekkora lesz a napi hőtermelés

Különböző modellek segítségével megvizsgáljuk, hogy a felállított hipotézisek valamelyike elvethető-e, vagy elfogadható-e. Illetve egy év hány napján és mekkora pontossággal lehet a 9:45 és 10:00 között rögzített adatokból előre jelezni, hogy milyen idő lesz aznap. Mivel a 9:45 és 10:00 között összegyűjtött adatokból próbáljuk megjósolni a 0-24 órás időtartamot, ezért a másodpercenként rögzített adatokat elő kell készíteni. A rendszerből txt kiterjesztéssel rendelkező fájlokat lehet letölteni. Egy fájl egy nap másodpercenként rögzített adatait tartalmazza. Ezen txt fájlokat be kell olvasni RStudio-ba, úgyhogy közben egy for ciklussal összegezni kell őket. Tehát egy napra minden egyes fizikai jellemzőből egy összegzett adat maradjon. A bemeneti adatok a 09:45 és 10:00 között összegzett adatok lesznek, így a hőmérséklet, a napkollektorok összesített hőmennyisége, a kollektorok hőmennyisége külön, a passzívház oldalán egy függőleges Heatpipe vákuumcsöves és a passzívház tetején döntött szögben elhelyezett síkcsöves és síklemezes rendszerű napkollektorok adatai szétválasztva, a szélesség összesített értéke, a külső hőmérséklet összesítve, illetve ezen fizikai jellemzők átlagai. A kimeneti fizikai jellemző az összes, 24 óra alatt összegyűjtött hőmennyiség, tehát ezt a fizikai jellemzőt fogjuk előre jelezni.

Két R szkriptet -et kell készítenünk és ezeket exportálunk két külön csv fájlba az adatokat. Az egyik fájl a 09:45 és 10:00 között rögzített adatok összegzését és átlagolását fogja tartalmazni,

a másik pedig a 24 óra alatt összegyűjtött összegzett hőmennyiséget. A teljes napi adatok a melléklet 11. számú ábrán látható R-szkript-el kerültek beolvasásra. Az R-szkript a 09:45 és 10:00 közötti adatok beolvasásához a melléklet 12. számú ábrája alatt található. Ezután egy másik R- szkripttel beolvassuk (lásd melléklet 13. számú ábra) látható az előzőleg kiírt két csv fájlt. A 24 órás adatokat tartalmazó fájlból az összehőtermelés oszlopra van szükségünk, mint kimeneti változó. A 15 perc adatait tartalmazó fájlból az összes változó bemeneti változóként fogjuk felhasználni. Ezenkívül az adatokat szétválasztjuk a csillagászati évszakok szerint. Ezt úgy tudjuk megtenni, hogy a data frame-ben négy oszlopot képzünk, ezeket az oszlopokat az évszakok szerint nevezzük el, majd binárisan, azaz nullával és eggyel megjelöljük az egyes adatsorokat. Ez azt jelenti, hogy az adatsor annál az évszaknál kap egyes kategóriát, amelyikhez tartozik, a többi oszlopban nulla lesz. Összesen 14 változónk lesz, ebből 13 bemeneti és egy kimeneti. A kimeneti fizikai jellemző az összes, 24 óra alatt összegyűjtött hőmennyiség, tehát ezt a fizikai jellemzőt fogjuk megpróbálni előre jelezni. A következő lépésben lefuttatjuk a ggpairs parancsot. A GGally csomag ggpairs() függvénye lehetővé teszi, hogy egy scatterplot mátrixot készítsünk. Az egyes párok szórásdiagramja a diagram bal oldalán látható, a Pearson-féle korrelációs érték és szignifikancia pedig a jobb oldalon látható (lásd melléklet 4. ábra). A korrelációs együtthatókból az látható, hogy a 9:45-10:00 óráig összegyűjtött prediktor változók, képzett prediktor változók és a 24 órás adatok között egy olyan korrelációs érték van, ami erős kapcsolatra enged következtetni és néhány közepesen erős kapcsolatot is találunk (lásd 1. Táblázat). Fontos, hogy az adatokat train és tesztelési halmazokra osszuk, amikor felügyelt tanulási algoritmusokat használunk, például lineáris regressziót, Random Forest-et, Naïve Bayes osztályozást, logisztikus regressziót és döntési fákat. A train adathalmazzal betanítjuk a gépi tanulási algoritmust, a teszt adathalmazzal ellenőrizzük. Jelen esetben 75% train-re és 25% tesz részre osztjuk az adathalmazunkat. Ezek, után lefuttatjuk a szkripteket mindegyik modellünkénél.

A következő modelleket állítottunk össze és vetjük össze egymással:

	Változók												
	Prediktorok 09:45-10:00												Predikció 0-24
	SumQp	SumTkulso	sumQ1	sumQ2	sumSzelseb	atlQ1	atlQ2	atlTkulso	atlSzelseb	Tavas	Nyar	Osz	Tel
Pearson féle korrelációs együttható	0.841	0.162	0.62	0.407	0.087	0.158	0.618	0.404	0.08	-0.19	0.245	-0.04	-0.02
Egyváltozós lineáris regresszió	✓												✓
Kétváltozós modell kereszthatás nélkül	✓		✓										✓
Kétváltozós modell kereszthatással	✓		✓										✓
Többváltozós lineáris modell	✓		✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
Neurális háló 1. 2 prediktor változóval													✓
Neurális háló 2. 11 prediktor változóval	✓		✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
													p-érték
													<0,05
													<0,05
													<0,05
													<0,05
													<0,05

Táblázat 1 A vizsgált modellek és változók

Forrás: Saját táblázat

SumQp	Összesített hőmennyiség
SumTkulso	Összesített hőmérséklet
sumQ1	Összesített hőmennyiség döntött kollektor
sumQ2	Összesített hőmennyiség függőleges kollektor
sumSzelseb	Összesített szélsébség
atlQ1	Átlagolt hőmennyiség
atlQ2	Átlagolt hőmennyiség
atlSzelseb	Átlagolt szélsébség
Tavas	Bináris adat
Nyar	Bináris adat
Osz	Bináris adat
Tel	Bináris adat

Táblázat 2 A változók rövidítéseinek a jelentése

Forrás: Saját táblázat

Szignifikanciaszint, vagy alfaszint $\alpha=0,05$, ez annak a valószínűsége, hogy az észlelt hatás a véletlen műve, illetve a kritériumszint, aminél a p szignifikancia értéknek kisebbnek kell lennie ahhoz, hogy elvethessük a nullhipotézist. Az összegzés p-értéke azt a nullhipotézist teszteli, hogy az együttható nulla (nincs hatás). Az alacsony p-érték ($\alpha < 0,05$) azt jelzi, hogy a nullhipotézist el lehet vetni. Más szóval, egy alacsony p-értékkel rendelkező prediktor valószínűleg hatásos tagja a modellnek, mivel a prediktor értékének változásai kapcsolatban állnak a válaszváltozó változásával. Ezzel szemben egy nagyobb (nem szignifikáns) p-érték arra utal, hogy a prediktorban bekövetkező változások nem állnak kapcsolatban a válaszban bekövetkező változásokkal. Ahogyan a 1. táblázatban is látható a regressziós modellek összegzésének lefuttatása 0,05 -nél kisebb p-értéket eredményezett. Így a H_0 hipotézisünket elvethetjük, tehát reggel 9:45 és 10:00 között beérkezett adatokból megjósolható, hogy mekkora lesz a napi hőtermelés.

Ezek után előre jelezzük a különböző modellek segítségével a napi összmennyiséggel, majd összevetjük a teszt adatbázisunk adataival és megvizsgáljuk, hogy a valós értékek mennyire térnek el az előrejelzéstől.

Az eredmények összefoglalása a 3. Táblázatban láthatóak, az eredményeket százzal megszoroztuk és felkerekítettük a jobb olvashatóság és összehasonlíthatóság érdekében.

		Az eltérések (%) Átlaga	Az eltérések (%) Mediánja	Az eltérések (%) szórása	Az eltérések (%) standard hibája	Becslés vs. Teszt df napok száma 5% eltérésen belül
Modell	Egyváltozós lineáris regresszió	491900.6	11	2216208	254216.5	5
	Kétváltozós modell kereszthatás nélkül	596095.2	10.5	2702553	310004.1	7
	Kétváltozós modell kereszthatással	546983.8	15	2478197	284268.6	9
	Többváltozós lineáris modell	499648.5	11	2195079	251792.9	9
	Neurális háló 1. 2 prediktor változóval	263840.9	11	1079060	123776.7	11
	Neurális háló 2. 11 prediktor változóval	254626.7	8.5	1794804	205878.1	11

Táblázat 3 Az előre jelzett napi értékek és a valóságban rögzített értékek eltérései.

Forrás: Saját táblázat

A táblázatból kivehető, hogy a legjobb eredményt a neurális háló használata produkálta, itt 5% eltérésen belül 11 nap hőmennyiségét (kW/h) előre tudtuk jelezni. A modellek vizuális összehasonlítása (diagrammok) a melléklet 10. ábráján láthatóak.

6. Összefoglalás

A gépi tanulás során nagyon sok releváns adatra van szükség, hogy mennyire az mindig a feladat komplexitásától függ. Az ökölszabály az az, hogy ahány szabadsági fokunk van (degree of freedom), tehát jelen esetben ahány napot meg szeretnénk jósolni, annyiszor 10 adatra van szükségünk, tehát ténylegesen használható előrejelzéshez 10 évi adatnak kellene rendelkezésre állnia (Dorfman, 2022). Ehhez képest a majdnem egy éves adatokból 11 napot 5% eltéréssel előre jelezni neurális hálóval véleményem szerint olyan eredménynek számít, ami azt mutatja, hogy több adat rendelkezésre állása esetén valószínűleg elfogadható és felhasználható eredményt kapnánk. Ezt támasztja alá az is, hogy a korrelációs mátrixunkban találunk egy erős és több közepesen erős korrelációs kapcsolatot is.

Irodalomjegyzék

Anil Jain, M., 2016. *www.ibm.com*. [Online]

Available at: <https://www.ibm.com/blogs/watson-health/the-5-vs-of-big-data/>

[Hozzáférés dátuma: 5 November 2022].

Bucioglu, N., 2021. <https://www.computerweekly.com/>. [Online]

Available at: <https://www.computerweekly.com/de/meinung/Deep-Learning-in-Storage-Systemen-ein-Etikettenschwindel>

[Hozzáférés dátuma: 5 November 2022].

COSO IT, 2017. *www.cosoit.com*. [Online]

Available at: <https://www.cosoit.com/explanation-of-3v-model-of-big-data>

[Hozzáférés dátuma: 5 November 2022].

Dorfman, E., 2022. *postindustria.com*. [Online]

Available at: <https://postindustria.com/how-much-data-is-required-for-machine-learning/>

[Hozzáférés dátuma: 6 November 2022].

dr.Némethné Gál, A., 2000. *Statisztika I-II..* Tatabánya: TRI-MESTER.

Litzel, N. & Dipl.-Ing. (FH) Luber, S., 2018. *www.bigdata-insider.de*. [Online]

Available at: <https://www.bigdata-insider.de/was-ist-r-a-707966/>

[Hozzáférés dátuma: 5 November 2022].

Microsoft, 2021. *Microsoft.com*. [Online]

Available at: <https://download.microsoft.com/download/3/5/b/35bb997f-a8c7-485d-8c56-19444dafd757/azure-machine-learning-algorithm-cheat-sheet-july-2021.pdf>

[Hozzáférés dátuma: 5 November 2022].

Microsoft, 2022. *azure.microsoft.com*. [Online]

Available at: <https://azure.microsoft.com/hu-hu/overview/what-is-artificial-intelligence/#how>

[Hozzáférés dátuma: 5 November 2022].

Molnár, L., 2021. *Az R alapjai, bevezetés a felügyelt gépi tanulásba. : .*

Molnár, L., 2022. *1evNap24ig.R R-Szkript Projektfeladat-oktatási anyag*. Tatabánya: EDUTUS EGYETEM.

Molnár, L., 2022. *1evNap945ig.R R-szkript Projektfeladat-oktatási anyag*. Tatabánya: EDUTUS EGYETEM.

Molnár, L., 2022. *DoryHipotezis.R R-szkript Projektfeladat-oktatási anyag*. Tatabánya: EDUTUS EGYETEM.

Neutelings, I., 2021. *tikz.net*. [Online]

Available at: https://tikz.net/wp-content/uploads/2021/12/neural_networks-001.png

[Hozzáférés dátuma: 5 November 2022].

Pumhösel, A., 2019. *www.derstandard.at*. [Online]

Available at: <https://www.derstandard.at/consent/tcf/story/2000097169579/ki-forscher-algorithmen-werden-uns-verrueckt-machen>

[Hozzáférés dátuma: 5 November 2022].

SPSSABC.HU, 2021. *spssabc.hu*. [Online]

Available at: <https://spssabc.hu/tobbvaltozos-elemzes/regresszio/nem-linearis-regresszio/>

[Hozzáférés dátuma: 5 November 2022].

TIBCO Software Inc., 2022. *www.tibco.com*. [Online]

Available at: <https://www.tibco.com/de/reference-center/what-is-r-analytics>

[Hozzáférés dátuma: 5 November 2022].

Vogels, D. W., 2018. *www.wiwo.de*. [Online]

Available at: <https://www.wiwo.de/unternehmen/mittelstand/hannovermesse/digitalisierung-die-erweiterung-der-digitalen-welt/21268454-3.html>

[Hozzáférés dátuma: 5 November 2022].

Wuttke, L., 2022. *datasolut.com*. [Online]

Available at: <https://datasolut.com/was-ist-machine-learning/>

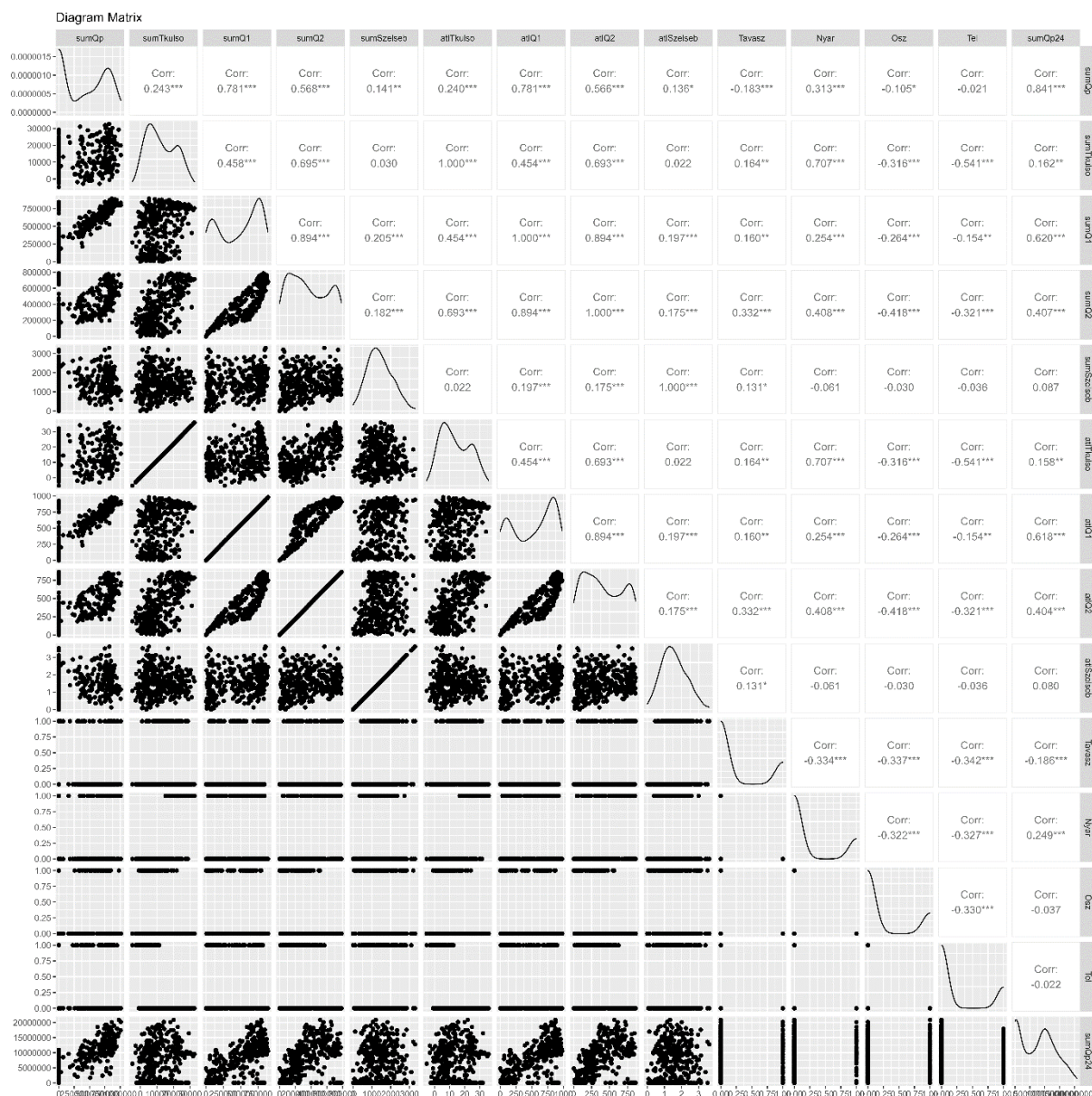
[Hozzáférés dátuma: 5 November 2022].

Wuttke, L., 2022. *datasolut.com*. [Online]

Available at: <https://datasolut.com/machine-learning-vs-deep-learning/>

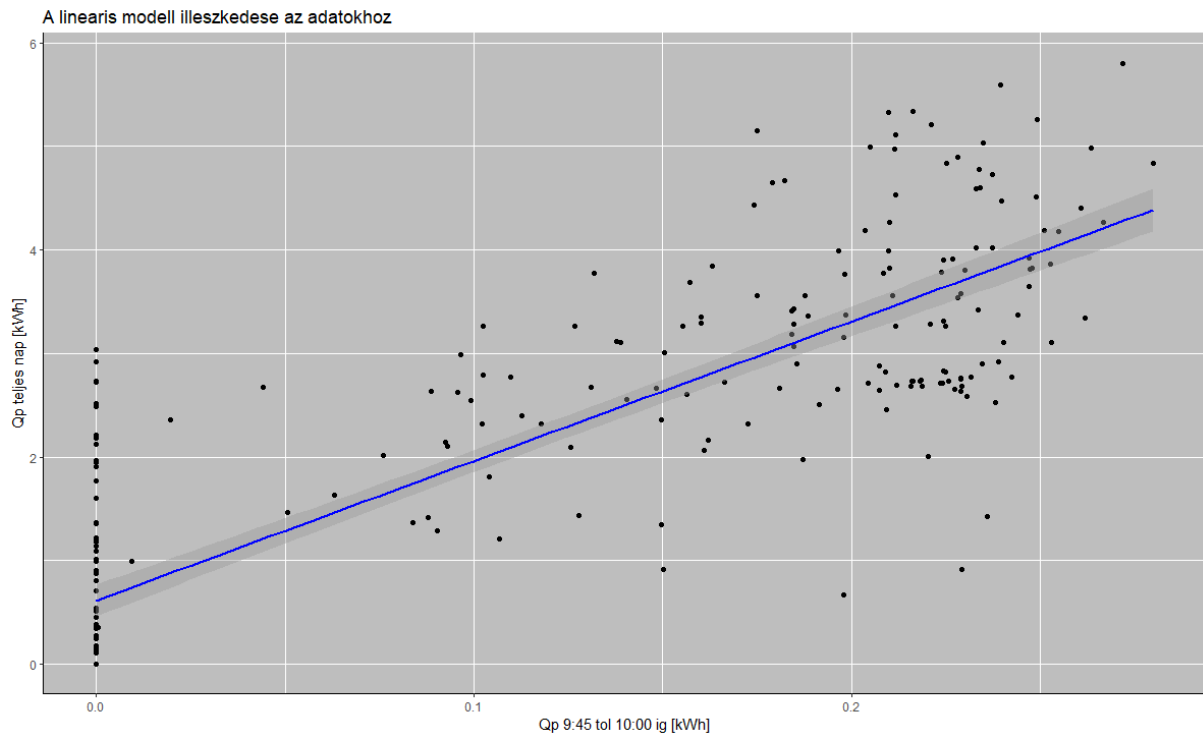
[Hozzáférés dátuma: 5 November 2022].

Mellékletek



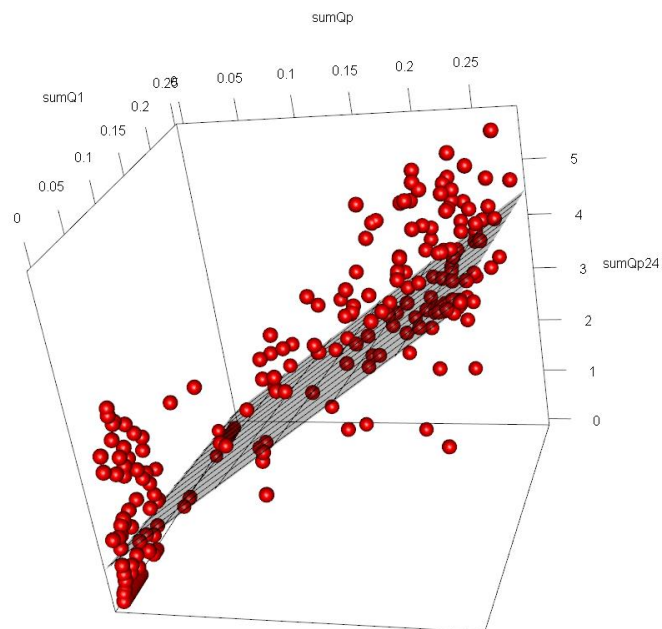
Ábra 4 Diagram mátrix

Forrás: saját ábra



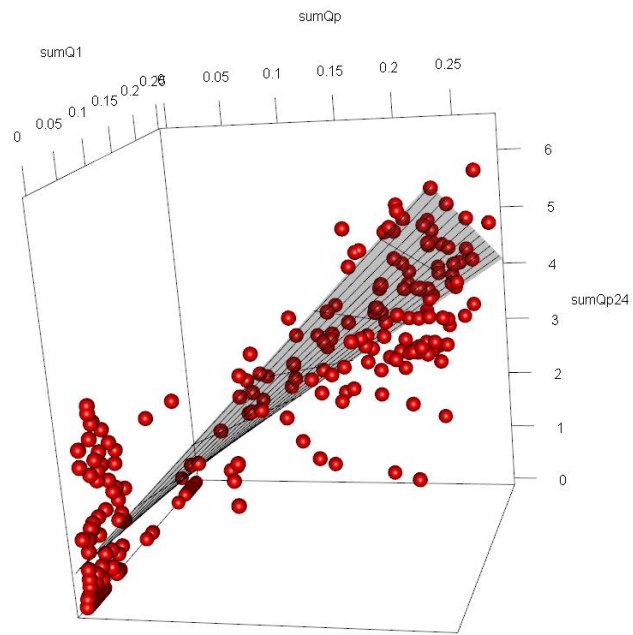
Ábra 5 Egyváltozós lineáris korreláció modell

Forrás: saját ábra



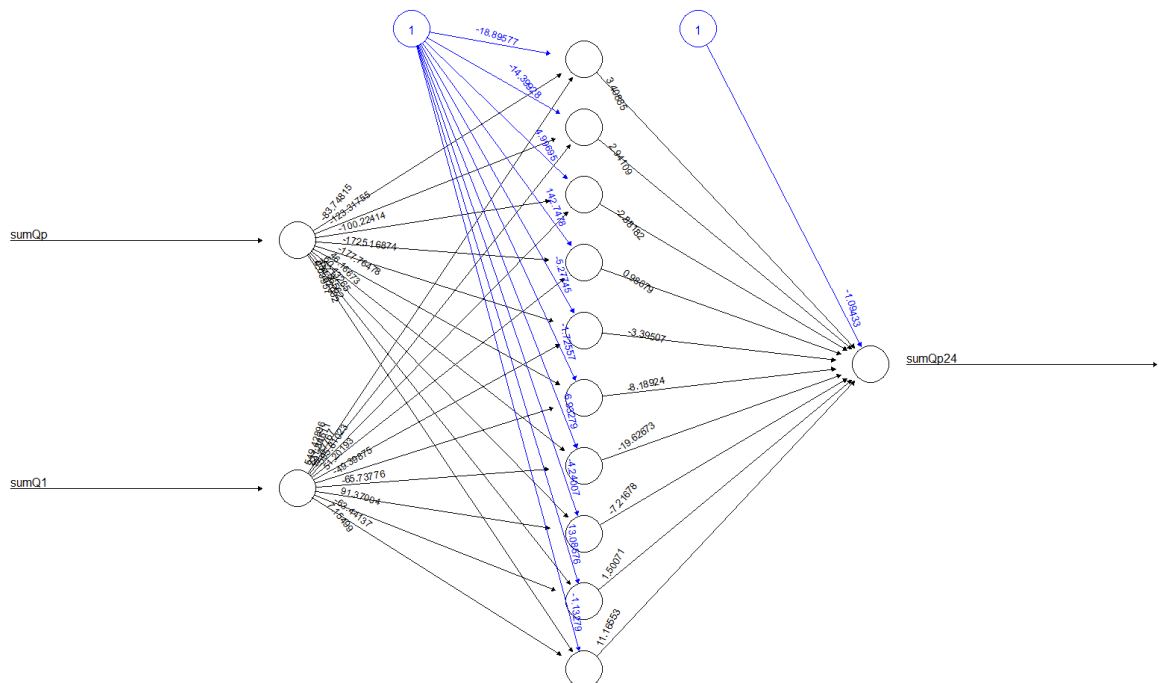
Ábra 6 Kétváltozós lineáris modell

Forrás: saját ábra



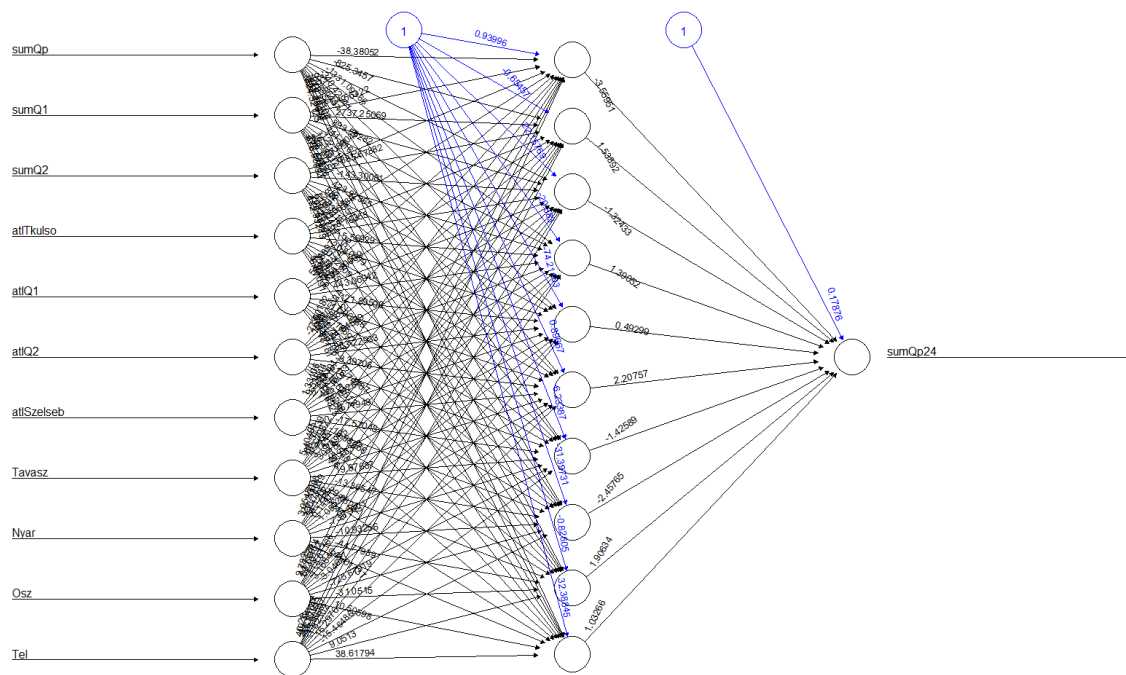
Ábra 7 Kétváltozós lineáris modell keresztthatással

Forrás: saját ábra



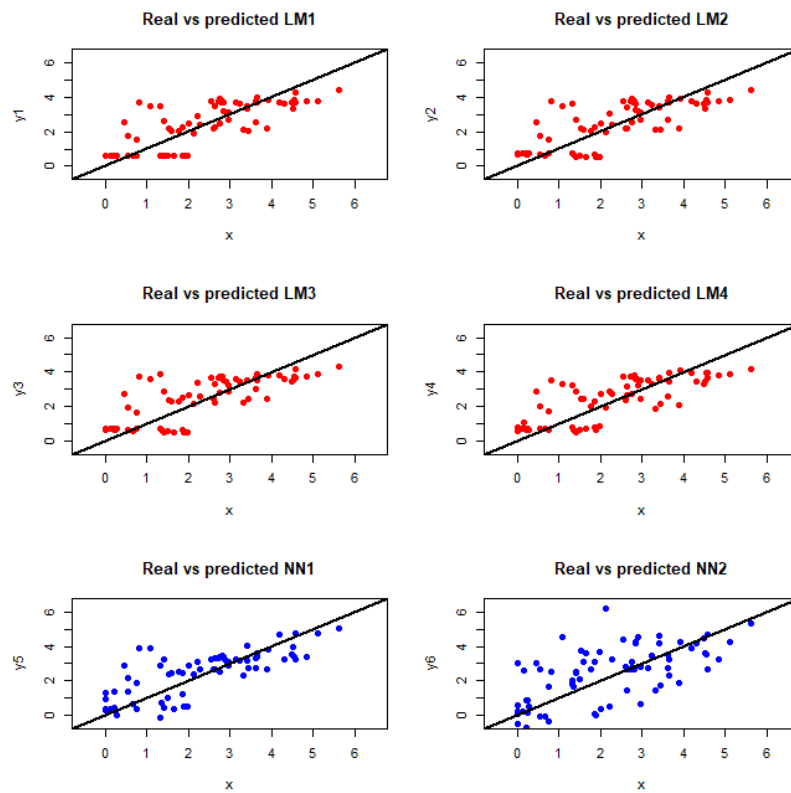
Ábra 8 Neurális hálózat tíz rejtett réteggel 2 prediktor változóval

Forrás: saját ábra



Ábra 9 Neurális hálózat tíz rejtett réteggel 12 prediktor változóval

Forrás: saját ábra



Ábra 10 A modellek összehasonlítása

Forrás: saját ábra

```
#szukseges csomagok meghivasa
```

```
library(tidyverse)
```

```
## Warning: Paket 'tidyverse' wurde unter R Version 4.1.3 erstellt
```

```
## -- Attaching packages ----- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr  1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
```

```
## Warning: Paket 'ggplot2' wurde unter R Version 4.1.3 erstellt
```

```
## Warning: Paket 'tibble' wurde unter R Version 4.1.3 erstellt
```

```
## Warning: Paket 'tidyr' wurde unter R Version 4.1.3 erstellt
```

```
## Warning: Paket 'readr' wurde unter R Version 4.1.3 erstellt
```

```
## Warning: Paket 'purrr' wurde unter R Version 4.1.3 erstellt
```

```
## Warning: Paket 'dplyr' wurde unter R Version 4.1.3 erstellt
```

```
## Warning: Paket 'stringr' wurde unter R Version 4.1.3 erstellt
```

```
## Warning: Paket 'forcats' wurde unter R Version 4.1.3 erstellt
```

```
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(fs)
```

```
## Warning: Paket 'fs' wurde unter R Version 4.1.3 erstellt
```

```
library(GGally)
```

```
## Warning: Paket 'GGally' wurde unter R Version 4.1.3 erstellt
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg      ggplot2
```

```
#a mappaban talalhato fajlok kilistazasa
```

```
filepath <- fs::dir_ls("X:/TDK_/1ev")
```

```
## Lista hosszanak meghatarozasa
```

```
db_fajl <- length(filepath)
```

```
#data frame létrehozasa
```

```
napiMeres <- data.frame(matrix(ncol=6,nrow=365-9))
```

```
napiMeres[is.na(napiMeres)] <- 0
```

```
napiMeres <- napiMeres %>% rename(Datum=X1,
                                sumQp=X2,
                                sumTkulso=X3,
                                sumQ1=X4,
                                sumQ2=X5,
                                sumSzelseb=X6)
```

```
#fajlok beolvasasa for ciklussal
```

```
for (i in 1:db_fajl) {
```

```
mereslist <- read.csv(filepath[i], sep="\t", header = F)
```

```
meres2 <- mereslist[c(1,3,87,137,139,141)]
```

```
meres2 <- meres2 %>% rename(Datum=V1,
```

```

        Qp=V3,
        Tkulso=V87,
        Q1=V137,
        Q2=V139,
        Szelseb=V141)

##formatumok rendezese

day <- as.character(as.Date(substring(filepath[i],15,22), format = "%Y%m%d"))
nap <- gsub("-", ".", day)
meres <- meres2[meres2$Datum >= nap,]

meres$Qp <- gsub(",", ".", meres$Qp)
meres$Tkulso <- gsub(",", ".", meres$Tkulso)
meres$Q1 <- gsub(",", ".", meres$Q1)
meres$Q2 <- gsub(",", ".", meres$Q2)
meres$Szelseb <- gsub(",", ".", meres$Szelseb)

meres$Qp <- as.numeric(meres$Qp)
meres$Tkulso <- as.numeric(meres$Tkulso)
meres$Q1 <- as.numeric(meres$Q1)
meres$Q2 <- as.numeric(meres$Q2)
meres$Szelseb <- as.numeric(meres$Szelseb)

##formatumok rendezese a hinyzo adatok "NA" helyettesitese 0-val

meres[is.na(meres)] <- 0

## Datumok indexelese a for ciklushoz

napiMeres$Datum[i] <- meres$Datum[1]

## uj valtozok létrehozasa atlagolassal es osszegelessel

napiMeres$sumQp[i] <- sum(meres$Qp)
napiMeres$sumTkulso[i] <- sum(meres$Tkulso)
napiMeres$sumQ1[i] <- sum(meres$Q1)
napiMeres$sumQ2[i] <- sum(meres$Q2)
napiMeres$sumSzelseb[i] <- sum(meres$Szelseb)

napiMeres$atlQp[i] <- mean(meres$Qp)
napiMeres$atlTkulso[i] <- mean(meres$Tkulso)
napiMeres$atlQ1[i] <- mean(meres$Q1)
napiMeres$atlQ2[i] <- mean(meres$Q2)
napiMeres$atlSzelseb[i] <- mean(meres$Szelseb)

##Indexek kiiratas a beolvasas soran

print(i)
}

## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
## [1] 6
## [1] 7
## [1] 8
## [1] 9
## [1] 10
## [1] 11
## [1] 12
## [1] 13
## [1] 14
## [1] 15
## [1] 16
## [1] 17
## [1] 18
## [1] 19
## [1] 20
## [1] 21
## [1] 22
## [1] 23
## [1] 24

```

```
## [1] 25
## [1] 26
## [1] 27
## [1] 28
## [1] 29
## [1] 30
## [1] 31
## [1] 32
## [1] 33
## [1] 34
## [1] 35
## [1] 36
## [1] 37
## [1] 38
## [1] 39
## [1] 40
## [1] 41
## [1] 42
## [1] 43
## [1] 44
## [1] 45
## [1] 46
## [1] 47
## [1] 48
## [1] 49
## [1] 50
## [1] 51
## [1] 52
## [1] 53
## [1] 54
## [1] 55
## [1] 56
## [1] 57
## [1] 58
## [1] 59
## [1] 60
## [1] 61
## [1] 62
## [1] 63
## [1] 64
## [1] 65
## [1] 66
## [1] 67
## [1] 68
## [1] 69
## [1] 70
## [1] 71
## [1] 72
## [1] 73
## [1] 74
## [1] 75
## [1] 76
## [1] 77
## [1] 78
## [1] 79
## [1] 80
## [1] 81
## [1] 82
## [1] 83
## [1] 84
## [1] 85
## [1] 86
## [1] 87
## [1] 88
## [1] 89
## [1] 90
## [1] 91
## [1] 92
## [1] 93
## [1] 94
## [1] 95
## [1] 96
## [1] 97
## [1] 98
## [1] 99
```

```
## [1] 100
## [1] 101
## [1] 102
## [1] 103
## [1] 104
## [1] 105
## [1] 106
## [1] 107
## [1] 108
## [1] 109
## [1] 110
## [1] 111
## [1] 112
## [1] 113
## [1] 114
## [1] 115
## [1] 116
## [1] 117
## [1] 118
## [1] 119
## [1] 120
## [1] 121
## [1] 122
## [1] 123
## [1] 124
## [1] 125
## [1] 126
## [1] 127
## [1] 128
## [1] 129
## [1] 130
## [1] 131
## [1] 132
## [1] 133
## [1] 134
## [1] 135
## [1] 136
## [1] 137
## [1] 138
## [1] 139
## [1] 140
## [1] 141
## [1] 142
## [1] 143
## [1] 144
## [1] 145
## [1] 146
## [1] 147
## [1] 148
## [1] 149
## [1] 150
## [1] 151
## [1] 152
## [1] 153
## [1] 154
## [1] 155
## [1] 156
## [1] 157
## [1] 158
## [1] 159
## [1] 160
## [1] 161
## [1] 162
## [1] 163
## [1] 164
## [1] 165
## [1] 166
## [1] 167
## [1] 168
## [1] 169
## [1] 170
## [1] 171
## [1] 172
## [1] 173
## [1] 174
```

```
## [1] 175
## [1] 176
## [1] 177
## [1] 178
## [1] 179
## [1] 180
## [1] 181
## [1] 182
## [1] 183
## [1] 184
## [1] 185
## [1] 186
## [1] 187
## [1] 188
## [1] 189
## [1] 190
## [1] 191
## [1] 192
## [1] 193
## [1] 194
## [1] 195
## [1] 196
## [1] 197
## [1] 198
## [1] 199
## [1] 200
## [1] 201
## [1] 202
## [1] 203
## [1] 204
## [1] 205
## [1] 206
## [1] 207
## [1] 208
## [1] 209
## [1] 210
## [1] 211
## [1] 212
## [1] 213
## [1] 214
## [1] 215
## [1] 216
## [1] 217
## [1] 218
## [1] 219
## [1] 220
## [1] 221
## [1] 222
## [1] 223
## [1] 224
## [1] 225
## [1] 226
## [1] 227
## [1] 228
## [1] 229
## [1] 230
## [1] 231
## [1] 232
## [1] 233
## [1] 234
## [1] 235
## [1] 236
## [1] 237
## [1] 238
## [1] 239
## [1] 240
## [1] 241
## [1] 242
## [1] 243
## [1] 244
## [1] 245
## [1] 246
## [1] 247
## [1] 248
## [1] 249
```

```
## [1] 250
## [1] 251
## [1] 252
## [1] 253
## [1] 254
## [1] 255
## [1] 256
## [1] 257
## [1] 258
## [1] 259
## [1] 260
## [1] 261
## [1] 262
## [1] 263
## [1] 264
## [1] 265
## [1] 266
## [1] 267
## [1] 268
## [1] 269
## [1] 270
## [1] 271
## [1] 272
## [1] 273
## [1] 274
## [1] 275
## [1] 276
## [1] 277
## [1] 278
## [1] 279
## [1] 280
## [1] 281
## [1] 282
## [1] 283
## [1] 284
## [1] 285
## [1] 286
## [1] 287
## [1] 288
## [1] 289
## [1] 290
## [1] 291
## [1] 292
## [1] 293
## [1] 294
## [1] 295
## [1] 296
## [1] 297
## [1] 298
## [1] 299
## [1] 300
## [1] 301
## [1] 302
## [1] 303
## [1] 304
## [1] 305
## [1] 306
## [1] 307
## [1] 308
## [1] 309
## [1] 310
## [1] 311
## [1] 312
## [1] 313
## [1] 314
## [1] 315
## [1] 316
## [1] 317
## [1] 318
## [1] 319
## [1] 320
## [1] 321
## [1] 322
## [1] 323
## [1] 324
```

```
## [1] 325
## [1] 326
## [1] 327
## [1] 328
## [1] 329
## [1] 330
## [1] 331
## [1] 332
## [1] 333
## [1] 334
## [1] 335
## [1] 336
## [1] 337
## [1] 338
## [1] 339
## [1] 340
## [1] 341
## [1] 342
## [1] 343
## [1] 344
## [1] 345
## [1] 346
## [1] 347
## [1] 348
## [1] 349
## [1] 350
## [1] 351
## [1] 352
## [1] 353
## [1] 354
## [1] 355
## [1] 356
```

##eredmenyek csv fajlba irasa

```
write.csv(napiMeres, file = "napi24Meres.csv")
```

Ábra 11 R szkript a 0-24 óra alatt rögzített adatok beolvasásához

(Molnár, 2022)

1evNap945ig.R

Gabor

2022-11-05

#szukseges csomagok meghivasa

```
library(tidyverse)
```

```
## Warning: Paket 'tidyverse' wurde unter R Version 4.1.3 erstellt
```

```
## -- Attaching packages ----- tidyverse 1.3.2 --
```

```
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr  1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
```

```
## Warning: Paket 'ggplot2' wurde unter R Version 4.1.3 erstellt
```

```
## Warning: Paket 'tibble' wurde unter R Version 4.1.3 erstellt
```

```
## Warning: Paket 'tidyr' wurde unter R Version 4.1.3 erstellt
```

```
## Warning: Paket 'readr' wurde unter R Version 4.1.3 erstellt
```

```
## Warning: Paket 'purrr' wurde unter R Version 4.1.3 erstellt
```

```
## Warning: Paket 'dplyr' wurde unter R Version 4.1.3 erstellt
```

```
## Warning: Paket 'stringr' wurde unter R Version 4.1.3 erstellt
```

```
## Warning: Paket 'forcats' wurde unter R Version 4.1.3 erstellt
```

```

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag() masks stats::lag()

library(fs)

## Warning: Paket 'fs' wurde unter R Version 4.1.3 erstellt

library(GGally)

## Warning: Paket 'GGally' wurde unter R Version 4.1.3 erstellt

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg ggplot2

#a mappaban talalhato fajlok kilistazasa

filepath <- fs::dir_ls("X:/TDK_/1ev")

## Lista hosszának meghatározása
db_fajl <- length(filepath)

#data frame létrehozása

napiMeres <- data.frame(matrix(ncol=6,nrow=365-9))
napiMeres[is.na(napiMeres)] <- 0
napiMeres <- napiMeres %>% rename(Datum=X1,
                                sumQp=X2,
                                sumTkulso=X3,
                                sumQ1=X4,
                                sumQ2=X5,
                                sumSzelseb=X6)

#fajlok beolvasása for ciklussal

for (i in 1:db_fajl) {

mereslist <- read.csv(filepath[i], sep="\t", header = F)

meres2 <- mereslist[c(1,3,87,137,139,141)]
meres2 <- meres2 %>% rename(Datum=V1,
                           Qp=V3,
                           Tkulso=V87,
                           Q1=V137,
                           Q2=V139,
                           Szelseb=V141)

##formatumok rendezése

day <- as.character(as.Date(substring(filepath[i],15,22), format = "%Y%m%d"))
nap <- gsub("-", ".", day)
meres <- meres2[meres2$Datum >= nap,]

meres$ID <- c(1:nrow(meres))

meres$Qp <- gsub(",", ".", meres$Qp)
meres$Tkulso <- gsub(",", ".", meres$Tkulso)
meres$Q1 <- gsub(",", ".", meres$Q1)
meres$Q2 <- gsub(",", ".", meres$Q2)
meres$Szelseb <- gsub(",", ".", meres$Szelseb)

meres$Qp <- as.numeric(meres$Qp)
meres$Tkulso <- as.numeric(meres$Tkulso)
meres$Q1 <- as.numeric(meres$Q1)
meres$Q2 <- as.numeric(meres$Q2)
meres$Szelseb <- as.numeric(meres$Szelseb)

##formatumok rendezése a hinyzo adatok "NA" helyettesítése 0-val
meres[is.na(meres)] <- 0

##9:45 es 10:00 kozott létrehozott sorok megtartása

```

```

meres <- meres[meres$ID > 30850 & meres$ID < 31761,]

## Datumok indexelese a for ciklushoz

napiMeres$Datum[i] <- meres$Datum[1]

## uj valtozok létrehozasa atlagolassal es osszegelessel

napiMeres$sumQp[i] <- sum(meres$Qp)
napiMeres$sumTkulso[i] <- sum(meres$Tkulso)
napiMeres$sumQ1[i] <- sum(meres$Q1)
napiMeres$sumQ2[i] <- sum(meres$Q2)
napiMeres$sumSzelseb[i] <- sum(meres$Szelseb)

napiMeres$atlQp[i] <- mean(meres$Qp)
napiMeres$atlTkulso[i] <- mean(meres$Tkulso)
napiMeres$atlQ1[i] <- mean(meres$Q1)
napiMeres$atlQ2[i] <- mean(meres$Q2)
napiMeres$atlSzelseb[i] <- mean(meres$Szelseb)

##Indexek kiiratasa a beolvasas soran

print(i)

}

## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
## [1] 6
## [1] 7
## [1] 8
## [1] 9
## [1] 10
## [1] 11
## [1] 12
## [1] 13
## [1] 14
## [1] 15
## [1] 16
## [1] 17
## [1] 18
## [1] 19
## [1] 20
## [1] 21
## [1] 22
## [1] 23
## [1] 24
## [1] 25
## [1] 26
## [1] 27
## [1] 28
## [1] 29
## [1] 30
## [1] 31
## [1] 32
## [1] 33
## [1] 34
## [1] 35
## [1] 36
## [1] 37
## [1] 38
## [1] 39
## [1] 40
## [1] 41
## [1] 42
## [1] 43
## [1] 44
## [1] 45
## [1] 46
## [1] 47
## [1] 48
## [1] 49

```

```
## [1] 50
## [1] 51
## [1] 52
## [1] 53
## [1] 54
## [1] 55
## [1] 56
## [1] 57
## [1] 58
## [1] 59
## [1] 60
## [1] 61
## [1] 62
## [1] 63
## [1] 64
## [1] 65
## [1] 66
## [1] 67
## [1] 68
## [1] 69
## [1] 70
## [1] 71
## [1] 72
## [1] 73
## [1] 74
## [1] 75
## [1] 76
## [1] 77
## [1] 78
## [1] 79
## [1] 80
## [1] 81
## [1] 82
## [1] 83
## [1] 84
## [1] 85
## [1] 86
## [1] 87
## [1] 88
## [1] 89
## [1] 90
## [1] 91
## [1] 92
## [1] 93
## [1] 94
## [1] 95
## [1] 96
## [1] 97
## [1] 98
## [1] 99
## [1] 100
## [1] 101
## [1] 102
## [1] 103
## [1] 104
## [1] 105
## [1] 106
## [1] 107
## [1] 108
## [1] 109
## [1] 110
## [1] 111
## [1] 112
## [1] 113
## [1] 114
## [1] 115
## [1] 116
## [1] 117
## [1] 118
## [1] 119
## [1] 120
## [1] 121
## [1] 122
## [1] 123
## [1] 124
```

```
## [1] 125
## [1] 126
## [1] 127
## [1] 128
## [1] 129
## [1] 130
## [1] 131
## [1] 132
## [1] 133
## [1] 134
## [1] 135
## [1] 136
## [1] 137
## [1] 138
## [1] 139
## [1] 140
## [1] 141
## [1] 142
## [1] 143
## [1] 144
## [1] 145
## [1] 146
## [1] 147
## [1] 148
## [1] 149
## [1] 150
## [1] 151
## [1] 152
## [1] 153
## [1] 154
## [1] 155
## [1] 156
## [1] 157
## [1] 158
## [1] 159
## [1] 160
## [1] 161
## [1] 162
## [1] 163
## [1] 164
## [1] 165
## [1] 166
## [1] 167
## [1] 168
## [1] 169
## [1] 170
## [1] 171
## [1] 172
## [1] 173
## [1] 174
## [1] 175
## [1] 176
## [1] 177
## [1] 178
## [1] 179
## [1] 180
## [1] 181
## [1] 182
## [1] 183
## [1] 184
## [1] 185
## [1] 186
## [1] 187
## [1] 188
## [1] 189
## [1] 190
## [1] 191
## [1] 192
## [1] 193
## [1] 194
## [1] 195
## [1] 196
## [1] 197
## [1] 198
## [1] 199
```

```
## [1] 200
## [1] 201
## [1] 202
## [1] 203
## [1] 204
## [1] 205
## [1] 206
## [1] 207
## [1] 208
## [1] 209
## [1] 210
## [1] 211
## [1] 212
## [1] 213
## [1] 214
## [1] 215
## [1] 216
## [1] 217
## [1] 218
## [1] 219
## [1] 220
## [1] 221
## [1] 222
## [1] 223
## [1] 224
## [1] 225
## [1] 226
## [1] 227
## [1] 228
## [1] 229
## [1] 230
## [1] 231
## [1] 232
## [1] 233
## [1] 234
## [1] 235
## [1] 236
## [1] 237
## [1] 238
## [1] 239
## [1] 240
## [1] 241
## [1] 242
## [1] 243
## [1] 244
## [1] 245
## [1] 246
## [1] 247
## [1] 248
## [1] 249
## [1] 250
## [1] 251
## [1] 252
## [1] 253
## [1] 254
## [1] 255
## [1] 256
## [1] 257
## [1] 258
## [1] 259
## [1] 260
## [1] 261
## [1] 262
## [1] 263
## [1] 264
## [1] 265
## [1] 266
## [1] 267
## [1] 268
## [1] 269
## [1] 270
## [1] 271
## [1] 272
## [1] 273
## [1] 274
```

```
## [1] 275
## [1] 276
## [1] 277
## [1] 278
## [1] 279
## [1] 280
## [1] 281
## [1] 282
## [1] 283
## [1] 284
## [1] 285
## [1] 286
## [1] 287
## [1] 288
## [1] 289
## [1] 290
## [1] 291
## [1] 292
## [1] 293
## [1] 294
## [1] 295
## [1] 296
## [1] 297
## [1] 298
## [1] 299
## [1] 300
## [1] 301
## [1] 302
## [1] 303
## [1] 304
## [1] 305
## [1] 306
## [1] 307
## [1] 308
## [1] 309
## [1] 310
## [1] 311
## [1] 312
## [1] 313
## [1] 314
## [1] 315
## [1] 316
## [1] 317
## [1] 318
## [1] 319
## [1] 320
## [1] 321
## [1] 322
## [1] 323
## [1] 324
## [1] 325
## [1] 326
## [1] 327
## [1] 328
## [1] 329
## [1] 330
## [1] 331
## [1] 332
## [1] 333
## [1] 334
## [1] 335
## [1] 336
## [1] 337
## [1] 338
## [1] 339
## [1] 340
## [1] 341
## [1] 342
## [1] 343
## [1] 344
## [1] 345
## [1] 346
## [1] 347
## [1] 348
## [1] 349
```

```
## [1] 350
## [1] 351
## [1] 352
## [1] 353
## [1] 354
## [1] 355
## [1] 356
```

##eredmenyek csv fajlba irasa

```
write.csv(napiMeres, file = "napi945Meres.csv")
```

Ábra 12 R szkript a 9:45 - 10:00 között rögzített adatok beolvasásához

(Molnár, 2022)

Hipotezis.R

Gabor

2022-11-05

#szukseges csomagok meghivasa

```
library(tidyverse)
```

```
## Warning: Paket 'tidyverse' wurde unter R Version 4.1.3 erstellt
```

```
## -- Attaching packages ----- tidyverse 1.3.2 --
```

```
## v ggplot2 3.3.6      v purrr  0.3.5
## v tibble  3.1.8      v dplyr  1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
```

```
## Warning: Paket 'ggplot2' wurde unter R Version 4.1.3 erstellt
```

```
## Warning: Paket 'tibble' wurde unter R Version 4.1.3 erstellt
```

```
## Warning: Paket 'tidyr' wurde unter R Version 4.1.3 erstellt
```

```
## Warning: Paket 'readr' wurde unter R Version 4.1.3 erstellt
```

```
## Warning: Paket 'purrr' wurde unter R Version 4.1.3 erstellt
```

```
## Warning: Paket 'dplyr' wurde unter R Version 4.1.3 erstellt
```

```
## Warning: Paket 'stringr' wurde unter R Version 4.1.3 erstellt
```

```
## Warning: Paket 'forcats' wurde unter R Version 4.1.3 erstellt
```

```
## -- Conflicts ----- tidyverse_conflicts() --
```

```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(fs)
```

```
## Warning: Paket 'fs' wurde unter R Version 4.1.3 erstellt
```

```
library(GGally)
```

```
## Warning: Paket 'GGally' wurde unter R Version 4.1.3 erstellt
```

```
## Registered S3 method overwritten by 'GGally':
```

```
##   method from
##   +.gg ggplot2
```

```
library(knitr)
```

```
## Warning: Paket 'knitr' wurde unter R Version 4.1.3 erstellt
```

##Munkakönyvtár beallítása

```
setwd("H:/TDK")
```

###csv fajlok betöltése

```
napi945Meres <- read.csv("napi945Meres.csv")
napi24Meres <- read.csv("napi24Meres.csv")
```

```
#### a napi 9:45 - 10:00 kozott mert adatok beolvasasa hipotezis_df dataframebe
```

```
hypotezis_df <- napi945Meres
```

Az adatok szétválasztása csillagászati évszakok szerint

```
hypotezis_df$Tavaszi <- ifelse(hypotezis_df$Datum >= "2021.03.21." & hypotezis_df$Datum < "2021.06.21.",
1,0)
hypotezis_df$Nyari <- ifelse(hypotezis_df$Datum >= "2021.06.21." & hypotezis_df$Datum < "2021.09.22.",1,
0)
hypotezis_df$Oszi <- ifelse(hypotezis_df$Datum >= "2021.09.22." & hypotezis_df$Datum < "2021.12.21.",1,
0)
hypotezis_df$Tel <- ifelse(hypotezis_df$Datum > "2021.12.21.",1, ifelse(hypotezis_df$Datum < "2021.03.21.",1,0))
```

A teljes napi meresek napi összegzeset felvesszük oszlopként

```
hypotezis_df$sumQp24 <- napi24Meres$sumQp
```

A teljes napi merések napi összegzést felvesszük oszlopként

```
write.csv(hypotezis_df, "hypotezis_df.csv", row.names = FALSE)
```

```
korrdM <- ggpairs (data = hypotezis_df, columns = c(3:7, 9:17), title = "Diagram Matrix")
korrdM
```

```
## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removing 1 row that contained a missing value
```

```
## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removing 1 row that contained a missing value
```

```
## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removing 1 row that contained a missing value
```

```
## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removing 1 row that contained a missing value
```

```
## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removing 1 row that contained a missing value
```

```
## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removing 1 row that contained a missing value
```

```
## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removing 1 row that contained a missing value
```

```
## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removing 1 row that contained a missing value
```

```
## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removing 1 row that contained a missing value
```

```
## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removing 1 row that contained a missing value
```

```
## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removing 1 row that contained a missing value
```

```
## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removing 1 row that contained a missing value
```

```
## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removing 1 row that contained a missing value
```

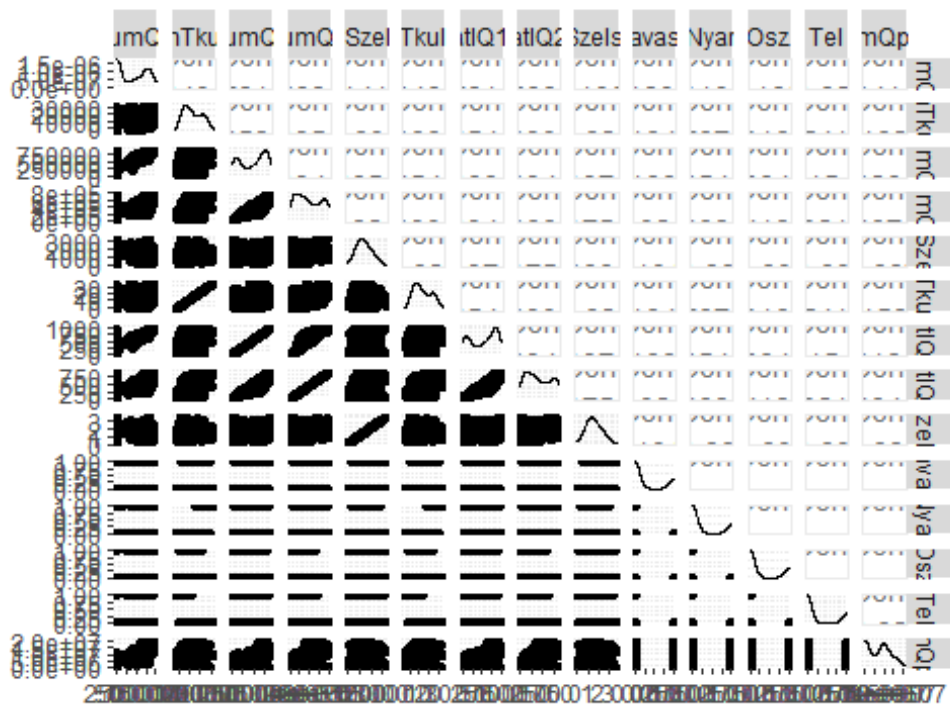
[illegible]

[illegible]

[illegible]

[illegible]

Diagram Matrix



```
# kiiras png kepfajlba a jobb olvashatosag erdekeben
## Limitsize= NULL jiapcsoljuk a nyomtathatosag felso hatarat

ggsave(filename = "hypotezis_df_DiagramMatrix.png",
        plot = korrDM, device = "png",
        path = "F:/TDK",
        height = 400,
        width = 400,
        units = "mm",
        dpi = 600,
        limitsize = 0)

## Warning in dir.create(dir, recursive = TRUE): kann Verzeichnis 'F:\TDK' nicht
## erzeugen, Grund 'Permission denied'

## Warning in normalizePath(path.expand(path), winslash, mustWork): path[1]="F:/
## TDK": Das Gerät ist nicht bereit

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removing 1 row that contained a missing value

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removing 1 row that contained a missing value

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removing 1 row that contained a missing value

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removing 1 row that contained a missing value

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removing 1 row that contained a missing value

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removing 1 row that contained a missing value

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removing 1 row that contained a missing value

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removing 1 row that contained a missing value

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
```


[illegible]

[illegible]

```

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removing 1 row that contained a missing value

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removing 1 row that contained a missing value

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removing 1 row that contained a missing value

## Warning: Removed 1 rows containing missing values (geom_point).
## Removed 1 rows containing missing values (geom_point).
## Removed 1 rows containing missing values (geom_point).
## Removed 1 rows containing missing values (geom_point).
## Removed 1 rows containing missing values (geom_point).
## Removed 1 rows containing missing values (geom_point).
## Removed 1 rows containing missing values (geom_point).
## Removed 1 rows containing missing values (geom_point).
## Removed 1 rows containing missing values (geom_point).

## Warning: Removed 1 rows containing non-finite values (stat_density).

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removing 1 row that contained a missing value

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removing 1 row that contained a missing value

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removing 1 row that contained a missing value

## Warning: Removed 1 rows containing missing values (geom_point).
## Removed 1 rows containing missing values (geom_point).
## Removed 1 rows containing missing values (geom_point).
## Removed 1 rows containing missing values (geom_point).
## Removed 1 rows containing missing values (geom_point).
## Removed 1 rows containing missing values (geom_point).
## Removed 1 rows containing missing values (geom_point).
## Removed 1 rows containing missing values (geom_point).
## Removed 1 rows containing missing values (geom_point).
## Removed 1 rows containing missing values (geom_point).

## Warning: Removed 1 rows containing non-finite values (stat_density).

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removing 1 row that contained a missing value

## Warning: Removed 1 rows containing missing values (geom_point).
## Removed 1 rows containing missing values (geom_point).
## Removed 1 rows containing missing values (geom_point).
## Removed 1 rows containing missing values (geom_point).

```

```

## Removed 1 rows containing missing values (geom_point).
## Removed 1 rows containing missing values (geom_point).
## Removed 1 rows containing missing values (geom_point).
## Removed 1 rows containing missing values (geom_point).

## Warning in grDevices::dev.off(): agg could not write to the given file

# Linearis regressziós modell

library(caret)

## Warning: Paket 'caret' wurde unter R Version 4.1.3 erstellt

## Lade nötiges Paket: lattice
##
## Attache Paket: 'caret'
##
## Das folgende Objekt ist maskiert 'package:purrr':
##
## lift

library(dplyr)
library(scatterplot3d)

## Warning: Paket 'scatterplot3d' wurde unter R Version 4.1.3 erstellt

library(rgl)

## Warning: Paket 'rgl' wurde unter R Version 4.1.3 erstellt

library(plotrix)

##
## Attache Paket: 'plotrix'
##
## Das folgende Objekt ist maskiert 'package:rgl':
##
## mtext3d

# A kovetkezo parancs kiterjeszti az eredmények konzolon megjelenitheto szamat.

options(scipen=999)

# A homennyiség oszlopok atvaltasa kW/h-ba

hypotezis_df$sumQp <- hypotezis_df$sumQp / (3600*1000)
hypotezis_df$atlQp <- hypotezis_df$atlQp / (3600*1000)
hypotezis_df$sumQ1 <- hypotezis_df$sumQ1 / (3600*1000)
hypotezis_df$atlQ1 <- hypotezis_df$atlQ1 / (3600*1000)
hypotezis_df$sumQ2 <- hypotezis_df$sumQ2 / (3600*1000)
hypotezis_df$atlQ2 <- hypotezis_df$atlQ2 / (3600*1000)
hypotezis_df$sumQp24 <- hypotezis_df$sumQp24 / (3600*1000)

### irányított veletlenség generalas

set.seed(42)

####Jelen esetben 25% train-re es 75% tesz reszre osztjuk az adathalmazunkat.

trainIndex <- createDataPartition (hypotezis_df$X, p = .75,
                                   list = FALSE, times = 1)

#####betanito es tesztelo data-framek létrehozasa
trainDF <- hypotezis_df[ trainIndex,]
testDF <- hypotezis_df[-trainIndex,]

# Modell 1: ebben a modelben az osszes homennyiség es az 9:45 es 10:00 kozotti homennyiseget vizsgaljuk
# linearis regreszsióval

```

```
## A summary file-al kilistazzuk a p<0,05 (szignifikancia)
### Nullhipotezist elvethetjuk

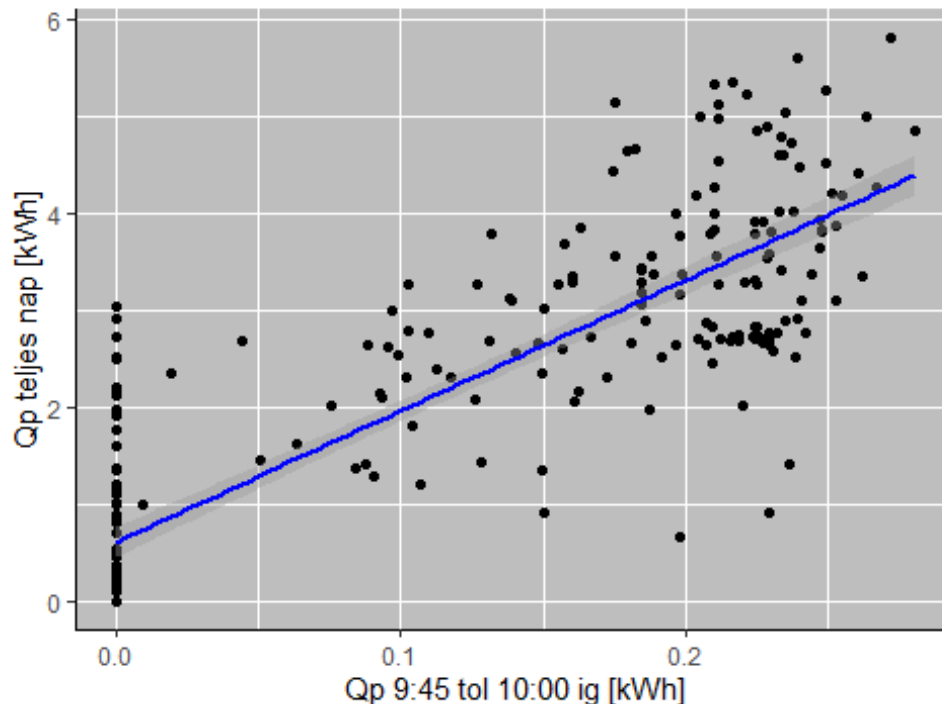
# Modell 1 -> predikcio y ertekek a 24 oras besugarzas, prediktor x ertekek a 9:45 -tol 10:00 -ig mert
besugarzas
fit_1 <- lm(sumQp24 ~ sumQp, data = trainDF)
summary(fit_1)

##
## Call:
## lm(formula = sumQp24 ~ sumQp, data = trainDF)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.7897 -0.6127 -0.2235  0.5610  2.4285
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)  0.61269    0.07819   7.836 0.000000000000112 ***
## sumQp       13.47747    0.52575  25.635 < 0.000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8852 on 266 degrees of freedom
## Multiple R-squared:  0.7119, Adjusted R-squared:  0.7108
## F-statistic: 657.1 on 1 and 266 DF, p-value: < 0.0000000000000022

ggplot(data = trainDF, aes(x = sumQp, y = sumQp24)) +
  geom_point() +
  stat_smooth(method = "lm", col = "blue") +
  theme(panel.background = element_rect(fill = "grey"),
        axis.line.x=element_line(),
        axis.line.y=element_line()) +
  ggtitle("A linearis modell illeszkedese az adatokhoz")+
  labs(x="Qp 9:45 tol 10:00 ig [kWh]", y="Qp teljes nap [kWh]")

## `geom_smooth()` using formula 'y ~ x'
```

A linearis modell illeszkedese az adatokhoz

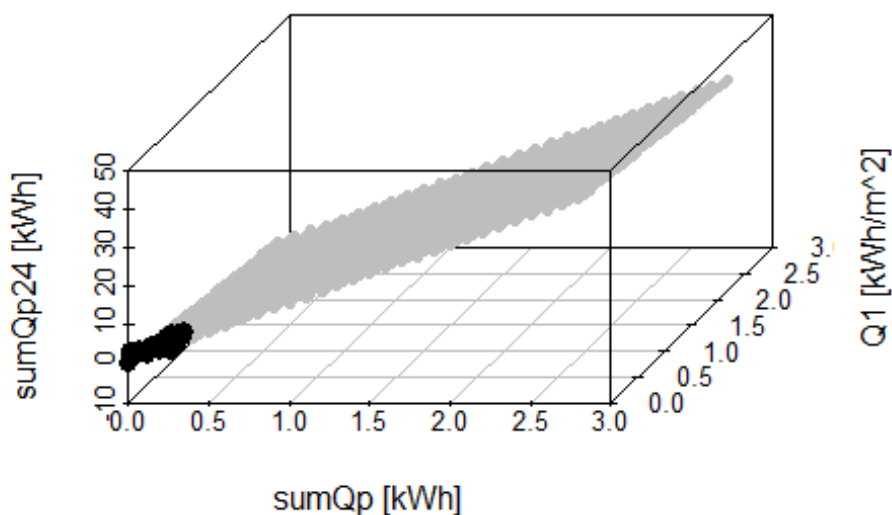


```
# Modell 2 -> három dimenzios abraulas a prediktor az osszes mert hoenergia, es a Q1 besugarzasmero z
tengely 9:45 -tol 10:00 -ig
fit_2 <- lm(sumQp24 ~ sumQp + sumQ1, data = trainDF)
summary(fit_2)
```

```
##
## Call:
## lm(formula = sumQp24 ~ sumQp + sumQ1, data = trainDF)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.9443 -0.7081 -0.2257  0.5308  2.3643
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)   0.7707     0.1051   7.331 0.000000000000277 ***
## sumQp         14.9568     0.8444  17.713 < 0.000000000000002 ***
## sumQ1          -2.3444     1.0520  -2.229    0.0267 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8786 on 265 degrees of freedom
## Multiple R-squared:  0.7172, Adjusted R-squared:  0.715
## F-statistic: 336 on 2 and 265 DF, p-value: < 0.0000000000000022

sumQp <- seq(0,2.8, by=0.1)
sumQ1 <- seq(0,2.8, by=0.1)
pred_grid <- expand.grid(sumQp = sumQp, sumQ1 = sumQ1)

pred_grid$sumQp24 <- predict(fit_2, new = pred_grid)
fit_2_sp <- scatterplot3d(pred_grid$sumQp,
  pred_grid$sumQ1,
  pred_grid$sumQp24,
  angle = 60,
  color = "grey",
  pch = 20,
  ylab = "Q1 [kWh/m^2]",
  xlab = "sumQp [kWh]",
  zlab = "sumQp24 [kWh]" )
fit_2_sp$points3d(trainDF$sumQp,
  trainDF$sumQ1,
  trainDF$sumQp24,
  pch=16,
  cex = 1.1)
```



```
### 3D abrazolas
open3d()
```

```

## wgl
## 1

plot3d(fit_2,
       which = 1, #1:3
       sharedMouse = FALSE,
       use_surface3d = TRUE,
       box = FALSE,
       type = "s",
       col = "red",
       radius = 0.075,
       plane.col = "grey",
       plane.alpha = 0.3,
       do_grid = TRUE,
       grid.col = "black",
       grid.alpha = 1,
       grid.steps = 10,
       sub.steps = 9)

# Modell 3 -> három dimenziós abrazolás a prediktor az összes mert hőenergia, és a Q1 besugarzasméretre z
# tengely 9:45 -tól 10: 00 -ig
## kereszthatással!
fit_3 <- lm(sumQp24 ~ sumQp * sumQ1, data = trainDF)
summary(fit_3)

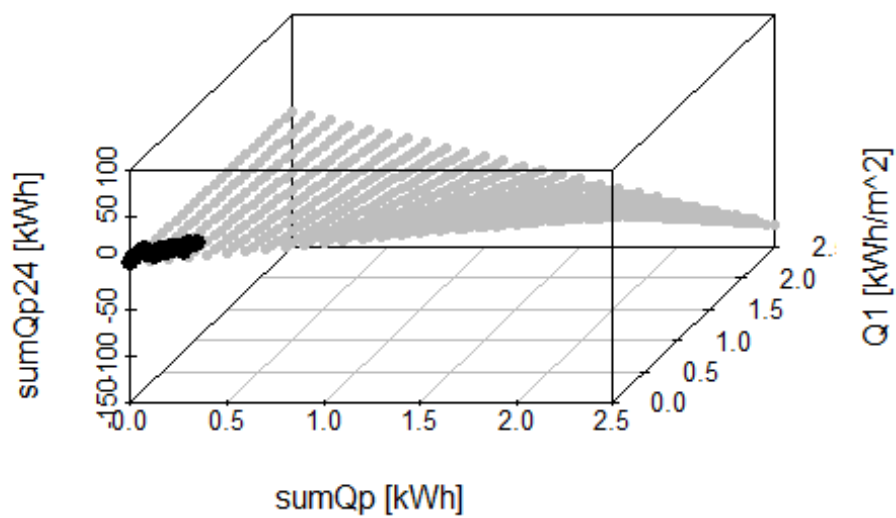
##
## Call:
## lm(formula = sumQp24 ~ sumQp * sumQ1, data = trainDF)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.2829 -0.6551 -0.2436  0.5326  2.4135
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)   0.7044     0.1087   6.479 0.00000000450014 ***
## sumQp         20.4801     2.6680   7.676 0.000000000000319 ***
## sumQ1         -1.9192     1.0627  -1.806   0.0721 .
## sumQp:sumQ1  -27.5779    12.6462  -2.181   0.0301 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8725 on 264 degrees of freedom
## Multiple R-squared:  0.7222, Adjusted R-squared:  0.719
## F-statistic: 228.7 on 3 and 264 DF, p-value: < 0.0000000000000022

sumQp <- seq(0,2.5, by=0.1)
sumQ1 <- seq(0,2.5, by=0.1)

pred_grid <- expand.grid(sumQp = sumQp, sumQ1 = sumQ1)

pred_grid$sumQp24 <- predict(fit_3, new = pred_grid)
fit_3_sp <- scatterplot3d(pred_grid$sumQp,
                        pred_grid$sumQ1,
                        pred_grid$sumQp24,
                        angle = 60,
                        color = "grey",
                        pch = 20,
                        ylab = "Q1 [kWh/m^2]",
                        xlab = "sumQp [kWh]",
                        zlab = "sumQp24 [kWh]" )
fit_3_sp$points3d(trainDF$sumQp,
                 trainDF$sumQ1,
                 trainDF$sumQp24,
                 pch=16,
                 cex = 1.1)

```



```
### 3D abrazolas
open3d()

## wgl
## 2

plot3d(fit_3,
  which = 1,
  sharedMouse = FALSE,
  use_surface3d = TRUE,
  box = FALSE,
  type = "s",
  col = "red",
  radius = 0.075,
  plane.col = "grey",
  plane.alpha = 0.3,
  do_grid = TRUE,
  grid.col = "black",
  grid.alpha = 1,
  grid.steps = 10,
  sub.steps = 9)

# Modell 4 --> a prediktor az osszes valtozo, predikcio valtoztatlanul az egesz napos mert besugarzas
fit_4 <- lm(sumQp24 ~ sumQp +
  sumQ1 +
  sumQ2 +
  atlTkulso +
  atlQ1 +
  atlQ2 +
  atlSzelseb +
  Tavasz +
  Nyar +
  Osz +
  Tel,
  data = trainDF)
summary(fit_4)

##
## Call:
## lm(formula = sumQp24 ~ sumQp + sumQ1 + sumQ2 + atlTkulso + atlQ1 +
##   atlQ2 + atlSzelseb + Tavasz + Nyar + Osz + Tel, data = trainDF)
##
## Residuals:
```

```
##      Min      1Q  Median      3Q      Max
## -3.1072 -0.6175 -0.1963  0.5205  2.3185
##
## Coefficients: (2 not defined because of singularities)
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)  0.498361   0.902111   0.552      0.5811
## sumQp       13.048365   1.137362  11.472 <0.0000000000000002 ***
## sumQ1        5.406191   2.859847   1.890    0.0598 .
## sumQ2       -8.882276   3.428263  -2.591    0.0101 *
## atlTkulso     0.002188   0.012346   0.177    0.8595
## atlQ1          NA          NA      NA         NA
## atlQ2          NA          NA      NA         NA
## atlSzelseb  -0.063441   0.076176  -0.833    0.4057
## Tavaszi      0.458314   0.927944   0.494    0.6218
## Nyar         0.621690   0.962537   0.646    0.5189
## Oszi         0.308481   0.905174   0.341    0.7335
## Tel         0.326286   0.899950   0.363    0.7172
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.875 on 257 degrees of freedom
## (1 Beobachtung als fehlend gelöscht)
## Multiple R-squared:  0.7263, Adjusted R-squared:  0.7167
## F-statistic: 75.78 on 9 and 257 DF, p-value: < 0.0000000000000022
```

```
# ##Neuralis halo
```

```
## Modell 5
```

```
library(neuralnet)
```

```
## Warning: Paket 'neuralnet' wurde unter R Version 4.1.3 erstellt
```

```
##
## Attache Paket: 'neuralnet'
##
## Das folgende Objekt ist maskiert 'package:dplyr':
##
##      compute
```

```
set.seed(42) ## a veletlenszam generator
```

```
## neuralis halozat 10 rejtett reteggel három dimenziós abraulas a prediktor az összes mert hoenergia,
## es a Q1 besugarzasmero z tengely 9:45 -tol 10: 00 -ig
## kereszthatassal!
```

```
trainDF[is.na(trainDF)] = 0
```

```
fit_5 <- neuralnet(sumQp24 ~ sumQp + sumQ1,
  data=trainDF,
  hidden=10, #c(6,3)
  act.fct = "logistic",
  linear.output = TRUE,
  stepmax = 1e7,
  # algorithm = "sag",
  startweights = NULL)
```

```
summary(fit_5)
```

```
##              Length Class      Mode
## call              8    -none-    call
## response         268    -none-  numeric
## covariate        536    -none-  numeric
## model.list         2    -none-    list
## err.fct            1    -none-  function
## act.fct            1    -none-  function
## linear.output      1    -none-  logical
## data              17  data.frame list
## exclude           0    -none-    NULL
## net.result         1    -none-    list
## weights            1    -none-    list
## generalized.weights 1    -none-    list
## startweights       1    -none-    list
## result.matrix     44    -none-  numeric
```

```

plot(fit_5)

set.seed(42)
fit_6 <- neuralnet(sumQp24 ~ sumQp +
  sumQ1 +
  sumQ2 +
  atlTkulso +
  atlQ1 +
  atlQ2 +
  atlSzelseb +
  Tavaszi +
  Nyar +
  Osz +
  Tel,
  data=trainDF,
  hidden=10, #c(6,3)
  act.fct = "logistic",
  linear.output = TRUE,
  stepmax = 1e7,
  # algorithm = "sag",
  startweights = NULL)

summary(fit_6)

##               Length Class      Mode
## call              8  -none-      call
## response          268  -none-    numeric
## covariate         2948  -none-    numeric
## model.list         2  -none-      list
## err.fct            1  -none-    function
## act.fct            1  -none-    function
## linear.output      1  -none-    logical
## data              17 data.frame list
## exclude           0  -none-      NULL
## net.result         1  -none-      list
## weights            1  -none-      list
## generalized.weights 1  -none-      list
## startweights       1  -none-      list
## result.matrix     134  -none-    numeric

plot(fit_6)

# Predikciók összevetése a teszt adatbázissal

test_eredm <- data.frame(ID = testDF$X)
test_eredm$sumQp24 <- testDF$sumQp24

pred1 <- predict(fit_1, testDF)
test_eredm$pred1 <- pred1

pred2 <- predict(fit_2, testDF)
test_eredm$pred2 <- pred2

pred3 <- predict(fit_3, testDF)
test_eredm$pred3 <- pred3

pred4 <- predict(fit_4, testDF)

## Warning in predict.lm(fit_4, testDF): prediction from a rank-deficient fit may
## be misleading

test_eredm$pred4 <- pred4

pred5 <- compute(fit_5, testDF)
test_eredm$pred5 <- pred5$net.result

pred6 <- compute(fit_6, testDF)
test_eredm$pred6 <- pred6$net.result

##A tesz eredmények felkerekítése

test_eredm$hiba1 <- round(100*(test_eredm$pred1 - test_eredm$sumQp24)
  / test_eredm$sumQp24, digits = 0)

```

```

test_eredm$hiba2 <- round(100*(test_eredm$pred2 - test_eredm$sumQp24)
/ test_eredm$sumQp24, digits = 0)

test_eredm$hiba3 <- round(100*(test_eredm$pred3 - test_eredm$sumQp24)
/ test_eredm$sumQp24, digits = 0)

test_eredm$hiba4 <- round(100*(test_eredm$pred4 - test_eredm$sumQp24)
/ test_eredm$sumQp24, digits = 0)

test_eredm$hiba5 <- round(100*(test_eredm$pred5 - test_eredm$sumQp24)
/ test_eredm$sumQp24, digits = 0)

test_eredm$hiba6 <- round(100*(test_eredm$pred6 - test_eredm$sumQp24)
/ test_eredm$sumQp24, digits = 0)

## data frame létrehozása ahol a becslesek a valós a teszt adatbázis valós adataival összevethetők

test_eredm_mod <- test_eredm[is.finite(rowSums(test_eredm)),]

## hibák átlagai

mean(test_eredm_mod$hiba1)

## [1] 491900.6

mean(test_eredm_mod$hiba2)

## [1] 596095.2

mean(test_eredm_mod$hiba3)

## [1] 546983.8

mean(test_eredm_mod$hiba4)

## [1] 499648.5

mean(test_eredm_mod$hiba5)

## [1] 263840.9

mean(test_eredm_mod$hiba6)

## [1] 254626.7

## hibák medianjai

median(test_eredm_mod$hiba1)

## [1] 11

median(test_eredm_mod$hiba2)

## [1] 10.5

median(test_eredm_mod$hiba3)

## [1] 15

median(test_eredm_mod$hiba4)

## [1] 11

median(test_eredm_mod$hiba5)

## [1] 11

median(test_eredm_mod$hiba6)

## [1] 8.5

```

```

## hibak szorasa

sd(test_eredm_mod$hiba1)
## [1] 2216208

sd(test_eredm_mod$hiba2)
## [1] 2702553

sd(test_eredm_mod$hiba3)
## [1] 2478197

sd(test_eredm_mod$hiba4)
## [1] 2195079

sd(test_eredm_mod$hiba5)
## [1] 1079060

sd(test_eredm_mod$hiba6)
## [1] 1794804

## standard hiba

std.error(test_eredm_mod$hiba1)
## [1] 254216.5

std.error(test_eredm_mod$hiba2)
## [1] 310004.1

std.error(test_eredm_mod$hiba3)
## [1] 284268.6

std.error(test_eredm_mod$hiba4)
## [1] 251792.9

std.error(test_eredm_mod$hiba5)
## [1] 123776.7

std.error(test_eredm_mod$hiba6)
## [1] 205878.1

# a megengedett elterés 5 kW/ h
megfeleloseg <- 5

# a napok 1 -el jelölése ahol a becsült és a jóslt elterés 5 kW/ h belül van

test_eredm_mod$megf1 <- ifelse(abs(test_eredm_mod$hiba1) <= megfeleloseg,1,0)
test_eredm_mod$megf2 <- ifelse(abs(test_eredm_mod$hiba2) <= megfeleloseg,1,0)
test_eredm_mod$megf3 <- ifelse(abs(test_eredm_mod$hiba3) <= megfeleloseg,1,0)
test_eredm_mod$megf4 <- ifelse(abs(test_eredm_mod$hiba4) <= megfeleloseg,1,0)
test_eredm_mod$megf5 <- ifelse(abs(test_eredm_mod$hiba5) <= megfeleloseg,1,0)
test_eredm_mod$megf6 <- ifelse(abs(test_eredm_mod$hiba6) <= megfeleloseg,1,0)

## Azon napok száma ahol meg lehetett jóslni 5 kW/h-n belüli elteréssel a napi homénnyiséget

sum(test_eredm_mod$megf1)
## [1] 5

sum(test_eredm_mod$megf2)

```

```

## [1] 7

sum(test_eredm_mod$megf3)

## [1] 9

sum(test_eredm_mod$megf4)

## [1] 9

sum(test_eredm_mod$megf5)

## [1] 11

sum(test_eredm_mod$megf6)

## [1] 11

# Real vs Predicted azaz a valos es az elorejelzett összehasonlítása
x <- test_eredm_mod$sumQp24
y1 <- test_eredm_mod$pred1
y2 <- test_eredm_mod$pred2
y3 <- test_eredm_mod$pred3
y4 <- test_eredm_mod$pred4
y5 <- test_eredm_mod$pred5
y6 <- test_eredm_mod$pred6

## a diagrammok egymás melle rendezese

par(mfrow=c(3,2))

## a diagrammok megjelenitese

plot(x,y1,col='red',main='Real vs predicted LM1',pch=19,cex=1, ylim=c(-0.5,6.5), xlim=c(-0.5,6.5))
abline(0,1,lwd=2)

plot(x,y2,col='red',main='Real vs predicted LM2',pch=19, cex=1, ylim=c(-0.5,6.5), xlim=c(-0.5,6.5))
abline(0,1,lwd=2)

plot(x,y3,col='red',main='Real vs predicted LM3',pch=19,cex=1, ylim=c(-0.5,6.5), xlim=c(-0.5,6.5))
abline(0,1,lwd=2)

plot(x,y4,col='red',main='Real vs predicted LM4',pch=19, cex=1, ylim=c(-0.5,6.5), xlim=c(-0.5,6.5))
abline(0,1,lwd=2)

plot(x,y5,col='blue',main='Real vs predicted NN1',pch=19,cex=1, ylim=c(-0.5,6.5), xlim=c(-0.5,6.5))
abline(0,1,lwd=2)

plot(x,y6,col='blue',main='Real vs predicted NN2',pch=19, cex=1, ylim=c(-0.5,6.5), xlim=c(-0.5,6.5))
abline(0,1,lwd=2)

```

Ábra 13 A különböző modellek összehasonlítása

(Molnár, 2022)