MAGYAR AGRÁR- ÉS ÉLETTUDOMÁNYI EGYETEM

A mesterséges intelligencia (MI) alkalmazási lehetőségei a szarvasmarha-tenyésztésben a küllemi bírálati eredmények, a szomatikus sejtszám és a kazeinmennyiség előrejelzésére

Doktori (PhD) értekezés

Tarr Bence

Gödöllő

(2025)

A doktori iskola megnevezése:	Műszaki Tudományi Doktori Iskola
Tudományága:	Agrárműszaki Tudományok
Vezetője:	Prof. Dr. Kalácska Gábor egyetemi tanár, DSc Magyar Agrár-és Élettudományi Egyetem, Műszaki Intézet
Témavezető:	Prof. Dr. Szabó István egyetemi tanár, PhD Magyar Agrár-és Élettudományi Egyetem, Műszaki Intézet Prof. Dr. Tőzsér János egyetemi tanár, az MTA doktora Széchenyi István Egyetem Albert Kázmér Mosonmagyaróvári Kar

. . . . . .

A témavezető jóváhagyása

. . . . . .

Az iskolavezető jóváhagyása

. . . . . .

A témavezető jóváhagyása

TARTALOMJEGYZÉK

JELÖLÉSJEGYZÉK

AI	Artificial Intelligence	Mesterséges intelligencia
MI	Mesterséges intelligencia	
ML	Machine Learning	Gépi tanulás
IoT	Internet of Things	
DL	Deep Learning	Mélytanulás
ANN	Artificial Neural Network	Mesterséges neurális hálók
UAV	Unmanned aerial vehicle	Pilóta nélküli légi jármű
SVM	Support Vector Model	
CNN	Convolutional Neural Network	Konvolúciós neurális háló
RFID	Radio Frequency IDentification	Rádiófrekvenciás azonosítás
LSTM	Long Short-Term Memory	
Mask R-CNN	Mask Region-based Convolutional Neural Network	
BCS	Body Condition Scoring	Testkondíció-becslés
DQN	Deep Q-Networks	
MSE	Mean Squared Error	Átlagos négyzetes hiba
SMOTE	Synthetic Minority Over-sampling Technique	
OLS	Ordinary Least Square	Legkisebb négyzetek módszere
FTIR-spektroszkópia	Fourier-transzformációs infravörös spektroszkópia	
R^2	Determinációs együttható	
$A\text{-}R^2$	Módosított determinációs együttható	
SCC	Somatic Cell Count	Szomatikus sejtszám
LSCC	Linearizált szomatikus sejtszám	

1. BEVEZETÉS

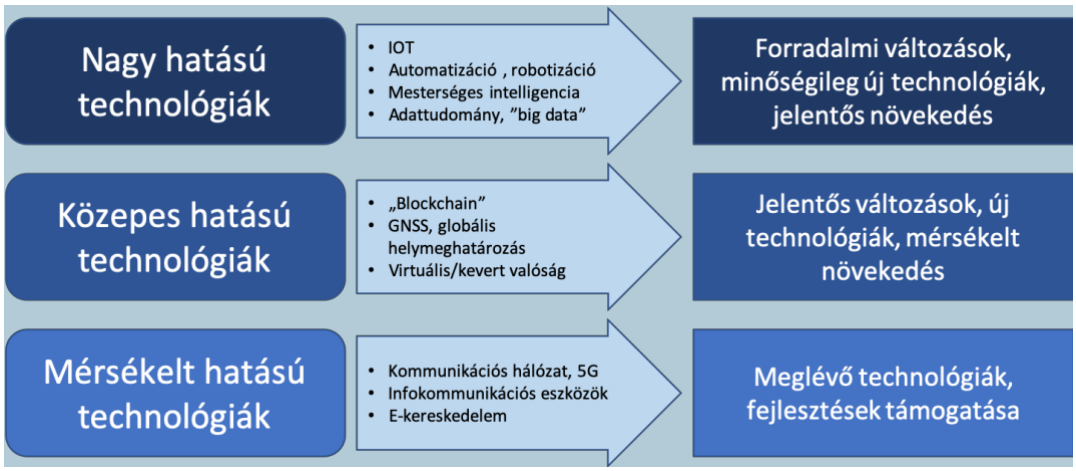
Több mint 25 évvel ezelőtt indult intenzívebb fejlődésnek a világban a precíziós gazdálkodás. Ezt az új elképzelést a mezőgazdaságban is alkalmazni kezdték, ugyanis ekkor már az informatikai és kommunikációs technológiák eléggé fejlettek, valamint elérhetőek voltak ehhez. 2017-ben *Khosia* már felhívja a figyelmet a precíziós gazdálkodás értelmezésének ellentmondásaira, nevezetesen: *„Érdekes megjegyezni, hogy a két évtizedes fejlődés után a helyspecifikus gazdálkodást sokan még mindig rosszul értelmezik, komplex és nagy technológiai innovációt értve a precíziós gazdálkodás alatt, miközben az ennél sokkal egyszerűbben is értelmezhető."*
Stafford (2017) véleménye szerint: *„A precíziós gazdálkodás nem egy új koncepció. Kijelenthetjük, hogy a jól gazdálkodó farmerek az évszázadok alatt olyan pontosak voltak, amit a technológia megengedett számukra".*
A precíziós gazdálkodás nem kívánja meg a nagy és összetett gépesítettséget. A lényeg az, hogy helyben milyen kreatív megoldásokat tudunk alkalmazni, milyen működést tudunk fenntartani és mindezt mennyire költséghatékonyan tudjuk tenni.
Khosia (2017) írásában azt is kifejti, hogy a technológia fejlődésével a precíziós gazdálkodás sem kerülheti el a nagy adatbázisok (*Big Data*) kezelését, használatát. Fontos utalni arra is, hogy napjainkban az olcsó szenzorok, a mindenütt megjelenő számítógépes rendszerek olyan mennyiségű adatot állítanak elő, amelynek feldolgozása és értelmezése hagyományos módszerekkel már nem elég hatékony. Ezek az adatok ráadásul meglehetősen olcsón állnak rendelkezésre, hiszen az adatok előállításhoz szükséges szenzorok, a továbbításhoz szükséges számítógépes hálózatok (vezetéknélküli hálózatok) ára folyamatosan csökken.

1.1. A precíziós gazdálkodás folyamata

Az adatgyűjtésen és az adatok értelmezésén alapuló, precíziós gazdálkodási mód több adatforrásból, részben automatikusan állít elő információt és végez beavatkozást, valamint támogatja a gazdaság működésével kapcsolatos döntési folyamatokat. Ez a technológia, egyebek mellett, abban is segít, hogy a gazdálkodási folyamat egyre több eleméről rendelkezzünk adattal, ami a folyamatok megértését és optimalizálását segítheti. A cél, hogy az adatgyűjtés és -felhasználás segítségével a hatékonyságot növelő, a gazdaság működését jobban megértő, a környezeti szempontokat is figyelembe vevő, hatékony gazdálkodás valósulhasson meg.

A precíziós gazdálkodás alapja az adatgyűjtés, amelyhez a hálózatban működő érzékelők, vagyis az IoT (Internet of Things) eszközök elterjedése teremtette meg a szükséges technikai alapot. A következő lépés az, hogy be kell gyűjteni és rendszerezni kell az ezen érzékelők által kapott adatokat. Az így rendelkezésre álló, rendezett adathalmazt alaposan elemezni, sőt szükség szerint validálni kell. Végül pedig vagy következtetéseket tudunk levonni az adatokból, vagy a jövőre nézve tudunk becsléseket készíteni, sőt felelős döntésket tudunk hozni arról, hogy hogyan érdemes beavatkozni a folyamatokba a kitűzött cél elérése érdekében.

Számos elemzés részletezi a különböző technológiák mezőgazdasági alkalmazásának hatását (*EP-AGRI,* 2019; *Szabó,* 2020). Az *1.1. ábrán* látható, hogy a legjelentősebb hatást a közeljövőben az IoT technológia és a mesterséges intelligencia alkalmazása jelentheti (a robotizáció és az adattudomány mellett).



1.1. ábra*:* Különböző IT-technológiák prognosztizált hatásténye zője

Az adatok elemzésében, feldolgozásában és a jövő tervezésében elengedhetetlen a mesterséges intelligencia (MI) jelentette technológiák használata.

1.2. A mesterséges intelligencia fogalma, elve, elemei

A mesterséges intelligenciának számos definíciója létezik. Általánosan elfogadott, hogy minden olyan gépet, vagy szoftvereszközt intelligensnek nevezünk, ami valamilyen cél érdekében racionálisan működik vagy cselekszik. Szerencsésebb, ha nem azt gondoljuk, hogy az intelligens rendszerek az emberi gondolkodást helyettesítik vagy kiváltják, hanem azt kiegészítik és ezzel kiterjesztik az emberi képességeket.

Egy másik megközelítés szerint intelligens tevékenységnek tarthatjuk a kép- és videófelismerést vagy a beszédértést is, illetve bizonyos kognitív képességek kialakítását. Vagyis a gép a „természetes intelligenciához" hasonlóan képes működni, döntéseket hozni. Az MI magasabb szintjein a rendszer szabályozottan képes viselkedésén változtatni, azaz „tanulni".

Számítástechnikai értelemben az MI a problémák algoritmusok segítségével történő intelligens megoldását jelenti. Nemcsak arról van tehát szó, hogy az algoritmusok „emberszerűen" viselkedjenek, hanem már az is egyfajta intelligenciának tekinthető, ha nagy, rendezetlen adathalmazokból nem egyértelmű eredményeket tudunk kinyerni.

Fontos látni tehát, hogy nemcsak az emberszerű gondolkodásról van szó, hanem egyszerűen már a gépi tanulás képessége is a mesterséges intelligencia körébe tartozik. Ez valami olyan eljárás, amely során a gép nagy adathalmazokban, önállóan képes mintákat felismerni úgy, hogy ezeket a mintákat, illetve ezek összefüggéseit matematikai algoritmusok alkalmazásával képes magától észrevenni és értelmezni. A tanulás vagy mélytanulás (az MI egyik ága) matematikai algoritmusok olyan összesége, amely által a gép képes olyan mintázatok felismerésére és értelmezésére, amire még az ember sem lenne képes – gépek segítsége nélkül. Ezeket a mintázatokat nagy adathalmazokon vagy akár objektumokon is értelmezni tudják (*Kollár et al*., 2021).

Az MI-nek tehát nincs egyértelmű és pontos meghatározása, többféle megközelítésben, többféle módon is leírhatjuk. Számos definíció létezik, szinte minden tudományág elkészítette a saját definícióit. *Russell és Norvig* (2005) szerint, ezek a definíciók két dimenzió mentén értelmezhetők. Az egyik értelmezés a gondolati folyamatokat és a következtetést célozza meg, míg a másik értelmezés a viselkedés felől közelíti meg a meghatározást. Nézhetjük azonban úgy is, hogy a helyes működést az emberi teljesítményhez mérjük, míg szintén elfogadható az intelligencia egy ideális koncepciója, amelyet mi racionalitásnak (rationality) nevezünk. Egy rendszer akkor racionális, ha valamilyen szempont vagy szabályrendszer alapján helyesen cselekszik.

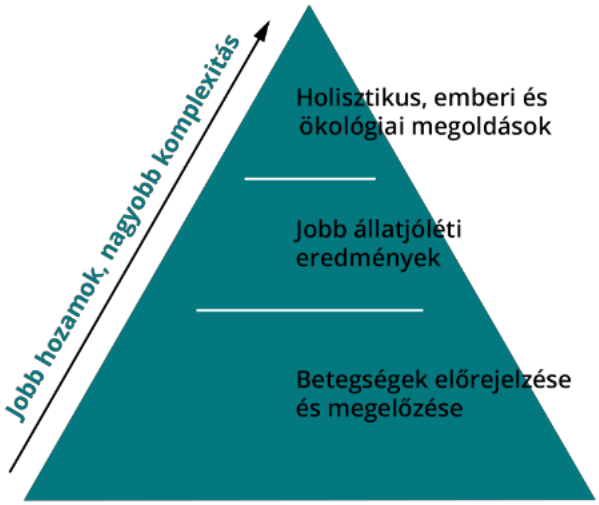
1.3. Az MI alkalmazásának tapasztalatai az állattenyésztésben

Az MI kezdetben egyetemi és vállalati kutatólaboratóriumok érdekes kísérleti fejlesztése volt, hiszen a nagyteljesítményű számítógépek és a bonyolult algoritmusok kezeléséhez szükséges tudás csak ezeken a helyeken volt elérhető. A technológia

fejlődésével, az MI-interfészek egyszerűsödésével és nem utolsó sorban a felhőszolgáltatások elterjedésével az MI szép lassan betört az élet minden területére. Ma szinte minden mobiltelefonban megtalálhatunk legalább egy olyan alkalmazást, ami mögött komoly gépi intelligencia működik, de az önvezető autók, traktorok és az egyre intelligensebb robotok is lassan megtalálják helyüket a mindennapi életben vagy az ipar területén.

Az IoT érzékelők és technológiák terjedése és széles körű használata a mezőgazdaságban is megalapozta az MI használatát. Hiszen a rendelkezésre álló jelentős adatmennyiség elemzése éppen az a terület, ahol az MI igazán segíteni tud.

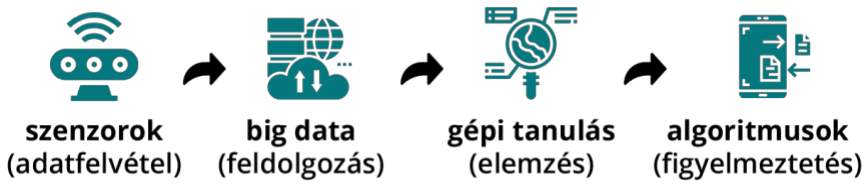
Suresh Neethirajan (2020) cikkében így összegzi a komplex adatokkal dolgozó állattenyésztő üzemek működését: „*A fejlett technológiák segítségével lehetőségünk lesz megérteni a komplex biológiai rendszerek működését. Segíthet az adatokból az igazán hasznos információk kinyerésében és ezáltal képesek leszünk az állattenyésztés bonyolult rendszerének megértésre*.” E rendszerek segíthetnek a kísérleti adatok gyűjtésében is, így pl. megmérhetjük a bendőben a lebomlási folyamatok frakcionált sebességét vagy az emlősejtek szaporodásának sebességét. A *1.2. ábrán* láthatjuk, hogyan javítja a fejlett technológia használata az állattenyésztés minőségi mutatóit.



1.2. ábra: Az állattenyésztés minőségi javításának hierarchiája, fejlett technológiák segítségével

Az állattenyésztés költséghatékonyságának legfontosabb mérőszáma az állománysűrűség, vagyis az, hogy mennyi állatot tudnak nevelni egy adott területen egységnyi idő alatt. Ezenkívül van még két nagyon fontos költségtényező: a takarmány költsége és a betegségek kezelési költségei. Mindkét költségtényező kezeléséhez és optimalizálásához jelenleg emberi közreműködésre, beavatkozásra van szükség. A

terület nagyságán kívül ez az egyik befolyásoló tényezője annak, hogy mennyire növelhető az állománysűrűség, vagyis, hogy mekkora kiegészítő munkaerőre van szükségünk az állatok ellátáshoz. Az MI-technológiák megjelenésével mind a takarmány összeállításban, mind a betegségmenedzsment terén számos segítő megoldással találkozhatunk. Végső soron tehát az MI-technológiák jelentősen csökkenthetik az állattenyésztés költségeit. Ennek oka pedig az okosérzékelők – az IoT-technológiák – terjedése az állattenyésztés legkülönbözőbb területein. Az okosérzékelők által szolgáltatott adatokat célszerű MI segítségével feldolgozni. A *1.3. ábrán* látható a szenzorokból érkező adatok feldolgozásának a folyamata.



1.3. ábra: Fejlett technológiák, amelyek szerepet játszanak az állattenyésztés során keletkező adatok felhasználásában

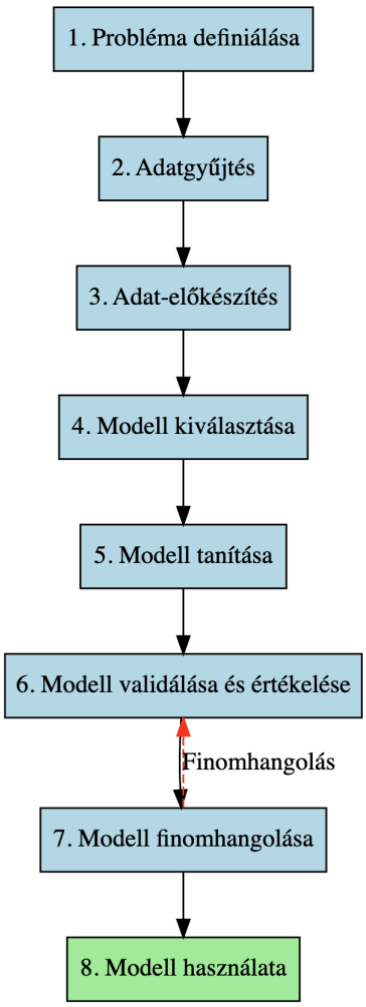
Egyszerű mágneses elmozdulásérzékelőkkel megfigyelhető többek között az állatok mozgásának sebessége, állapota. Kicsit bonyolultabb, GPS-alapú érzékelők pedig egészen pontos adatot tudnak szolgáltatni az állatok akár nagy területen való mozgásáról is. Kiegészítve ezeket hőmérsékletmérőkkel vagy egyéb kémiai vegyületeket érzékelő műszerekkel, máris összetett képet kaphatunk az állatok pillanatnyi állapotáról. A mesterséges intelligencia ezekből az adatokból jól felismerhető mintázatokat keresve figyelemmel tudja követni az állatok viselkedését, a viselkedési mintázatok változásából pedig következtetni lehet az állatok egészségére is. *Sadeghi et al.* (2015) egy tanulmányban azt is ismertették, hogy az MI előbb képes jelezni egy betegség kialakulását, mint a hagyományos emberi vizsgálatok. De számos más kutatás is megerősítette, hogy a Big Data és az MI-technológiák közösen hatékonyan képesek előre jelezni a betegségeket az állatállomány körében (*Vander Waal et al*., 2017; *Fernández-Carrión et al*., 2017).

A betegségek pontos előrejelzése nemcsak a költségeket fogja csökkenteni, hanem általa érezhetően csökkenthető a felhasznált antibiotikumok mennyisége is, amelynek nagy jelentősége van az emberek egészségének megőrzésének szempontjából.

Az emberek között is megfigyelhető jelenség az úgynevezett érzelmi átterjedés (emotional contagion). Ez azt jelenti, hogy az emberek követik a társaságukban lévő

másik ember érzelmi változásait, megnyilvánulásait és azt magukévá teszik. Az állatvilágban is megfigyelhető ez a jelenség, ahol ez az állatok gyors reagálását segíti elő például vészhelyzet esetén. Ennél fogva az állatok viselkedésének monitorozása és azonnali kiértékelése lehetővé teszi a teljes állomány hangulati, egészségi állapotának a követését és a szükséges beavatkozások időben történő meglépését.

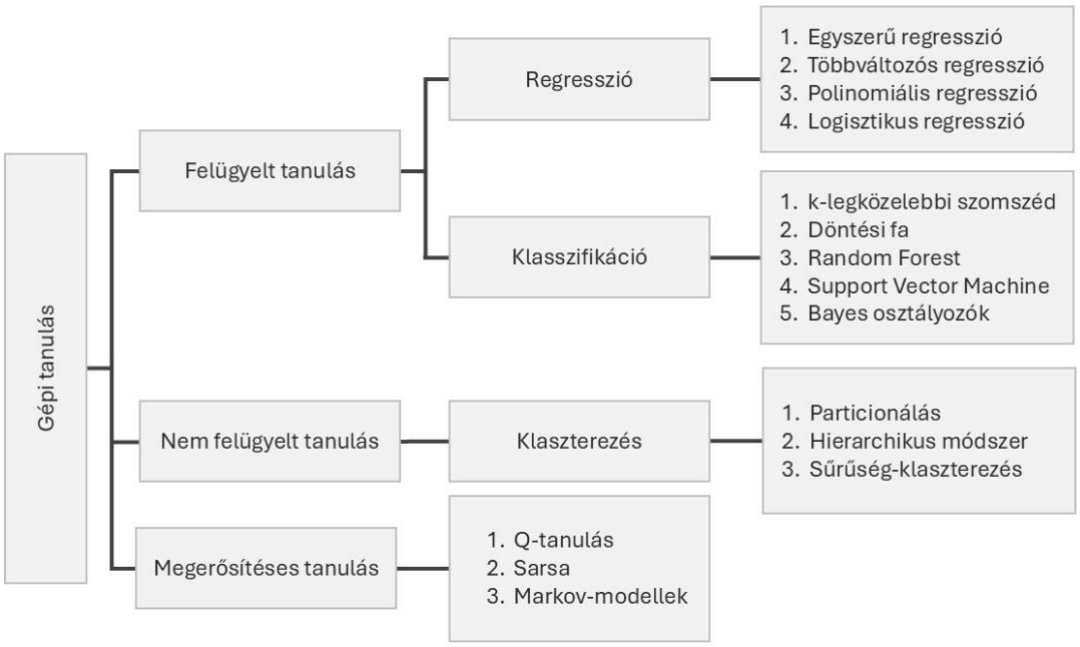
Az ML (az angol Machine Learning, azaz gépi tanulás kifejezés rövidítése) a mesterséges intelligencia egy részterülete. Olyan algoritmusokkal foglalkozik, amelyek képessé teszik a számítógépeket olyan feladatok végrehajtására, amire addig nem voltak felprogramozva. A gépi tanulás során egy szoftver az adatokból kinyeri az összefüggéseket és ez alapján készít egy algoritmust, amely alkalmassá teszi intelligens döntések meghozatalára. Az ML folyamat három kulcsfontosságú részből áll: adatbevitel, modellépítés és általánosítás *(1.4. ábra)*.



1.4. ábra: A gépi tanulás folyamata

Az általánosítás tulajdonképpen maga az előrejelzés, vagyis egy kimeneti érték meghatározása azon bemeneti adatokra, amelyekkel az algoritmus korábban nem volt betanítva.

Napjainkban a különböző ML-algoritmusok szinte mindenki számára hozzáférhetőek internetes erőforrásokon keresztül is, ezért használatukat az adatok szélesebb körére érdemes vizsgálni. Sőt, az erőforrások fejlődésével a mélytanulás (angolul Deep Learning, rövidítve DL) is elterjedt kutatási téma lett. A DL az ML-algoritmusok családjának az a csoportja, amely igazán nagy – sokszor strukturálatlan – adathalmazok elemzésére alkalmas. Ez a megoldás mesterséges neurális hálózatokat (ANN) használ az intelligens döntések meghozatalához. Az ML-algoritmusokat sokféleképpen lehet csoportosítani. Egy lehetséges felosztást az *1.5. ábrán* láthatunk.



1.5. ábra*:* Az ML-algoritmusok csoportosítása
(*Sharma et al*., 2021 alapján)

Elmondható, hogy az állattenyésztésben hatalmas felhalmozott adatvagyon áll rendelkezésre, valamint az is kijelenthető, hogy a gazdák, termelők, tenyésztők rutinszerűen gyűjtik és szabványos módon ellenőrzik a tenyésztés, termelés során keletkező adatokat.

2. CÉLKITŰZÉSEK

A mesterséges intelligencia használata az állattenyésztés területén is meghatározó feltétele a további mennyiségi és minőségi növekedésnek. A szakirodalom áttekintése után egyértelművé vált számomra, hogy a mesterséges intelligencia számos szolgáltatásának használata már megjelent a tenyésztők körében. Azonban a rendelkezésre álló „historikus adatok" feldolgozása, azokból becslő, előre jelző modellek készítése eddig nem kapott kellő tudományos hangsúlyt.

Ezért kutatásom tárgya, hogy a szarvasmarha-tenyésztésben fellelhető adatbázisokat megvizsgálva a tenyésztők számára hasznos becslő modelleket készítsek. Az adatok feldolgozására, statisztikai elemzésére eddig is történtek erőfeszítések, de ezek nem jelentek még meg a mindennapi használatban, és elsősorban hagyományos matematikai modelleket használtak.

Állat-egészségügyi és minőségbiztosítási okok miatt az állattenyésztésben és a tejtermelésben is hosszú évek óta kialakult rendje van az adatok gyűjtésének és elemzésének is, ezért nagy számú és jelentős méretű adatbázis, korábbi mérési, szakértői eredmény áll rendelkezésre.

A mesterséges intelligencia egyik leggyorsabban fejlődő ága, a gépi tanulás éppen ilyen adatbázisok felhasználásával képes meglepően pontos modellek, algoritmusok készítésére.

Célom tehát bizonyítani, hogy a meglévő adatok felhasználásával készíthetők olyan modellek, amelyek kellően pontos előrejelzést tudnak adni a szarvasmarha-tenyésztők számára bizonyos területeken.

Választásom kétféle konkrét területre esett: a szarvasmarhák küllemi bírálati pontozására, valamint a tejminőség ellenőrzésre. Mindkét szolgáltatásra nagy szükség van az állattenyésztésben, és különösen a küllemi bírálati pontozás esetében nagy a szubjektivitás kockázata.

Kutatásom fókuszában a szarvasmarha-tenyésztés és a tejtermelés során keletkező adatok feldolgozására, hasznosítására összpontosítottam. Kutatásom célja, hogy meglévő adatok feldolgozásával olyan, gépi tanuláson alapuló algoritmusokat, modelleket készítsek, amelyek segítik a termelés jobb tervezését, a tenyésztés támogatását. A kutatásaim során végül három probléma megoldásra kerestem MI-alapú megoldást: a szarvasmarhák küllemi bírálati pontszámainak becslésére, a tejben

lévő szomatikus sejtszám becslésére és a tejben lévő kazein mennyiségének becslésére.

A kutatáshoz nemcsak múltbéli adatok feldolgozása nyújt kellő bemeneti információt, hanem a modern termelőeszközök (pl. robot fejőgép) által automatikusan rögzített adatok is kiválóan alkalmasak ilyen modellek kialakításához.

Célkitűzésem tehát a rendelkezésre álló adatok feldolgozása után olyan, gépi tanuláson alapuló modellek keresése és teljesítményük ellenőrzése, melyek használhatók a tenyésztési folyamatok vagy a termelési folyamatok esetében előrejelzések támogatására.

Kutatásom támogatására olyan saját fejlesztésű szoftver rendszert tervezek fejleszteni, amely magában foglalja az adatfeldolgozás, adatelőkészítés, a tanító adatbázis kiválasztás, a tanítás és ellenőrzés minden lépést. Ez a rendszer jó alapja lehet a későbbiekben egy széleskörben is használható alkalmazás kifejlesztésének, amely nagyon kicsit beruházási igénnyel lenne használható a tenyésztők számára a fent említett témákban.

3. SZAKIRODALMI ÁTTEKINTÉS

A szakirodalom áttekintését a az egyik legjelentősebb tudományos portálon, a Web of Science-en kezdtem. Mivel kutatási témám a szarvasmarha-tenyésztés során felhasználható, gépi tanuláson alapuló modellek fejlesztésére fókuszál, így ebben a témában megjelent műveket tekintettem át. Alapvetően adatalapú megközelítéssel, meglévő adatok felhasználásával szeretnék foglalkozni, ezért ezt a paramétert is felhasználtam a releváns szakirodalmi kör beazonosításához.
A tudományos adatbázis feldolgozásakor a következő angol kifejezéseket alkalmaztam a szűrések során:

- artificial intelligence,
- machine learning,
- cattle,
- data.

Meglehetősen szűk találati listát kaptam, mindössze 152 dokumentum felelt meg az együttes keresési feltételeknek.
A szűrés eredménye a *3.1. ábrán* látható témakörökhöz tartozó cikkeket tartalmazta.

3.1. ábra: A Web of Science keresési eredménye témakörök szerint

Az eredményből látható, hogy a legtöbb cikk természetesen a mezőgazdasági tudományterületen jelent meg az említett kulcsszavakkal. Azonban a témaköröm

határterületi mivoltát jól mutatja, hogy számos más határos tudományterülethez is tartoznak cikkek a találati listáról.
Ez megfelel a valóságnak, hiszen a precíziós mezőgazdaság eszközei, módszerei és technológiái a különböző tudományterületek együttes eredményét használják fel a termelés, tenyésztés optimalizálásához.
A következőkben áttekintem a szakirodalomban fellelhető kutatások néhány fontos eredményét.

3.1. Gépi tanulás a növénytermesztésben

Gépi tanulás segítségével az elmúlt évtizedekben felhalmozott tudást és tapasztalatot tudjuk felhasználni a jövőbeli szakértői rendszerek tanításához. Ezzel nemcsak előjelzéseket készíthetünk vagy becsléseket tehetünk, hanem a rendszer folyamatos fejlesztésével biztosíthatjuk a modern technológiák előnyeinek beépülését a mezőgazdasági termelés tervezésébe.
Az előzőekben vegyes MI-megoldásokat mutattam be, a következőkben pedig áttekintem, hogy a mezőgazdaság mely területén használják már most is sikerrel a gépi tanulás módszereit.
A mezőgazdaság szerepe a következő időszakban is várhatóan egyre nőni fog, ahogy egyre több embert kell ellátni egyre kisebb termőterületen (*Fraser*, 2020). A szenzoradatok használatával a precíziós mezőgazdaság jobb termésátlagot képes biztosítani. Az állattenyésztő farmok pedig a szenzoradatok gyors feldolgozásával MI-alapú rendszerekké tudják fejleszteni saját gazdaságukat.

3.1.1. Terméshozam előrejelzése

A terméshozam-előrejelzés a precíziós mezőgazdaság egyik legfontosabb témája. Nagy jelentőségű a terméshozam feltérképezése, a termésbecslés, a terménykínálat és a kereslet összehangolása, a termelékenységet növelő növénytermesztés szempontjából. Kávészemek termésmennyiségének becslésére használja a gépi tanulás és képfeldolgozás módszereit *Ramos* egy 2017-ben megjelent munkájában (*Ramos et al.,* 2017). A módszer három kategóriába sorolja a kávétermést: betakarítható, nem betakarítható és figyelmen kívül hagyott érési stádiumú gyümölcsök. A képfeldolgozás után az algoritmus megbecsülte a kávégyümölcsök tömegét és érési fokát. A munka célja az volt, hogy információt nyújtson a kávétermelőknek a gazdasági haszon optimalizálásához és a mezőgazdasági munkák megtervezéséhez.

Egy másik, a terméshozam-előrejelzéshez használt tanulmányban, melyet *Amatya* és társai készítettek, egy gépi látáson alapuló rendszer kifejlesztését ismertették. A rendszer a cseresznye szüret közbeni rázását és elkapását végezte automatikusan *(Amatya et al.,* 2015*)*. A rendszer szegmentálja és felismeri a cseresznyét akkor is, ha takarásban van. A fejlesztés fő célja az volt, hogy csökkentse a kézi betakarítási és kezelési műveletek munkaerőigényét. *Sengupta et al.* (2014) egy korai termésértékelő rendszert ismertettek, amely éretlen zöld citrusfélék azonosítására szolgál szabadtéri körülmények között. A többi tanulmányhoz hasonlóan a kísérlet célja az volt, hogy a termesztők számára terméspecifikus információkat nyújtson, hogy segítse őket az ültetvényük optimalizálásában a nyereség és a termésnövelés szempontjából. *Ali et al.* (2017) egy mesterséges neurális hálót alkalmazó modellt mutattak be, amely multitemporális és távérzékelési adatokon alapult, a gyep biomassza-tartalmának (kg szárazanyag/ha/nap) becslésére. *Pantazi et al.* (2016) is a terméshozam-előrejelzéssel, konkrétan a búza terméshozamának előrejelzésével foglalkoztak. Az általuk kifejlesztett módszer műholdas felvételeket használt, és a pontosabb előrejelzés érdekében talajadatokkal korrigált növénynövekedési jellemzőket kapott. Pilóta nélküli légi járművel (UAV) rögzített távoli képek segítségével a paradicsom felismerésére szolgáló algoritmus is a gépi tanulás módszerét használja egy tanulmány szerint (*Senthilnath et al.,* 2016). *Su et al.* (2017) support-vector módszer segítségével a kínai időjárási állomásokról kapott alapvető földrajzi információk alapján dolgoztak ki módszert a rizs terméshozamának az előrejelzésére. *Kung et al*. (2016) kifejlesztettek egy általánosított módszert is a mezőgazdasági terméshozam előrejelzésére. Kutatásukban hosszú időszak adatait használták egy neurálisháló-alapú rendszer tanításához (1997–2014). A tanulmányban bemutatott rendszer regionális előrejelzésre használható (különösen Tajvanon) a mezőgazdasági termelőket támogatja, hogy elkerüljék a piaci kereslet és kínálat egyensúlytalanságát.

3.1.2. Betegség előrejelzése

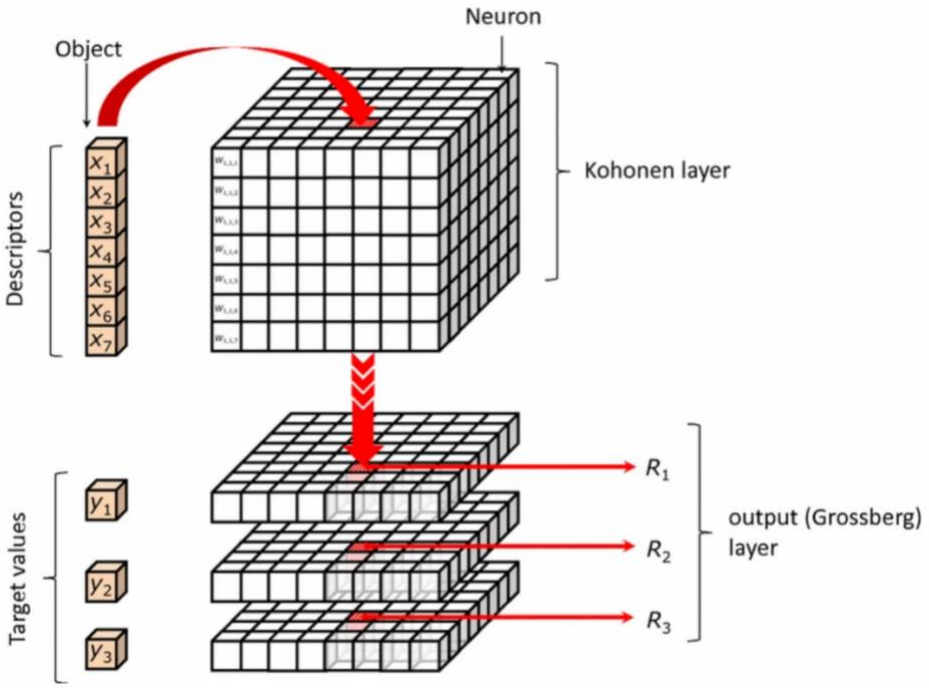
A betegségfelismerés és a terméshozam-előrejelzés ez egyik leggyakrabban előforduló témák a vonatkozó szakirodalomban. A mezőgazdaságban az egyik legjelentősebb probléma a kártevők és betegségek elleni védekezés szabadtéri (szántóföldi növénytermesztés) és üvegházi körülmények között. A kártevők és betegségek elleni védekezés legelterjedtebb gyakorlata a növényvédő szerek egyenletes permetezése a termőterületen. Ez a megoldás bár hatékony, magas pénzügyi és jelentős környezeti

költségekkel is jár. A környezeti hatások lehetnek többek között a növényi termékekben lévő szermaradványok, a talajvízszennyezés mellékhatásai, a helyi élővilágra és ökoszisztémákra gyakorolt hatások. Gépi tanuláson alapuló megoldások elsősorban az agrokémiai szerek pontosabb, célzott használatát segítik. *Pantazi et al.* (2017) cikkükben bemutatnak egy eszközt az egészséges és a *Microbotyum silybum* féregggombával fertőzött máriatövis *(Silybum marianum)* növények felismerésére. *Ebrahimi et al.* (2017*)* egy új, képfeldolgozási eljáráson alapuló módszert dolgoztak ki üvegházi környezetben nevelt szamócákon élősködő paraziták osztályozására és automatikus felismerésére. *Chung et al.* (2016) a rizspalántákban megtalálható Bakanae betegség kimutatására készítettek egy algoritmust. A kutatás célja a *Fusarium fujikuroi* kórokozó kimutatása volt a rizspalántákban. A fertőzött növények automatikus felismerése növelte a termés mennyiségét, és a szabad szemmel történő vizsgálathoz képest kevesebb időt vett igénybe.

A búza az egyik legfontosabb gabonaféle világszerte, alapvető élelmiszer-alapanyag. Az itt bemutatott tanulmányok mindegyike a beteg és egészséges búzanövények felismerésével és megkülönböztetésével foglalkozik. *Pantazi et al.* (2017) spektrális reflexiós képalkotási adatokon alapuló rendszert fejlesztettek ki a nitrogénstressz, valamint a sárgarozsdával fertőzött és az egészséges őszi búza megkülönböztetésére. A kutatás célja ezeknek a betegségeknek a pontos felismerése volt a gombaölő szerek és műtrágyák hatékonyabb, a növény igényeinek megfelelő felhasználása érdekében. *Moshou et al.* már 2014-ben bemutattak egy olyan rendszert, amely automatikusan megkülönböztette a vízstressznek kitett, *Septoria tritici* által fertőzött és az egészséges őszi búza lombkoronáját. Ez a megoldás optikai érzékelőket használt és a legkisebb négyzetek módszerét használó (LS)-SVM osztályozó algoritmus segítségével válogatta szét a képeket. *Moshou et al.* (2005–2006) további rendszereket mutattak be a sárgarozsdával fertőzött és az egészséges búza megkülönböztetésére. Ezek a rendszerek különböző matematikai és képfeldolgozó eljárások segítségével dolgoztak. *Ferentinos et al.* (2018) egy általánosabb, konvolúciós neurálisháló-alapú rendszert mutattak be a betegségek felismerésére egyszerű levélképek alapján, amely kellő pontossággal képes osztályozni az egészséges és beteg levelek között különböző növényeknél.

3.1.3. Gyomfelismerés

A mezőgazdaságban a gyomok felismerése és kezelése szintén jelentős probléma. Számos termelő a gyomokat jelöli meg a növénytermesztést fenyegető legfontosabb veszélyforrásként. A gyomok felismeréséhez szintén jól használhatók a különböző, gépi tanuláson alapuló módszerek. A gyomok automatikus felismerése lehetővé teszi a gyomok elpusztítására szolgáló eszközök, pl. robotok kifejlesztését, amelyek minimalizálják a gyomirtó szerek szükségességét. *Pantazi et al.* (2017) egy új módszert mutattak be, amely a szembeterjesztéses mesterséges neurális hálókat használja *(3.2. ábra)* (Counter Propagation (CP)-ANN), és a pilóta nélküli repülőgépek (UAS) által rögzített multispektrális képeken alapul a nehezen irtható és a terméshozamban jelentős veszteséget okozó gyomnövény, a *máriatövis* azonosítására.



3.2. ábra: A szembeterjesztéses neurális hálózat (CPANN) vázlata. Egy bemeneti objektum (X-értékek vektora – leíró értékek) leírókkal (balra fent) és végpontértékekkel (balra lent). A CPANN séma jobbra látható. A Kohonen-réteg van felül, a Kohonen-réteg alatt pedig a válaszfelületek megrajzolásához használt három válaszszintet tartalmazó kimeneti réteg található

Pantazi et al. (2016) egy másik tanulmányában egy új, ML technikákon és hiperspektrális képalkotáson alapuló módszert dolgoztak ki a növény- és gyomfajok felismerésére. A hiperspektrális képalkotás során számos (10–100-as nagyságrendű)

sávban készül spektrum az adott terület egyes pontjairól; az egyes spektrumokat magában foglaló képpontokból pedig kép állítható össze (képalkotó spektrometria).
A szerzők aktív tanulási rendszert hoztak létre a kukorica *(Zea mays)*, mint szántóföldi növényfaj, valamint a kúszó boglárka *(Ranunculus repens)*, mezei aszat *(Cirsium arvense)*, vadrepce *(Sinapis arvensis)*, tyúkhúr *(Stellaria media)*, gyermekláncfű *(Tarraxacum officinale)*, egynyári perje *(Poa annua)*, baracklevelű keserűfű *(Polygonum persicaria)*, nagy csalán *(Urtica dioica)*, felálló madársóska *(Oxalis europaea)* és komlós lucerna *(Medicago lupulina)* mint gyomfajok felismerésére. A fő cél e fajok pontos felismerése és megkülönböztetése volt gazdasági és környezetvédelmi célokból.

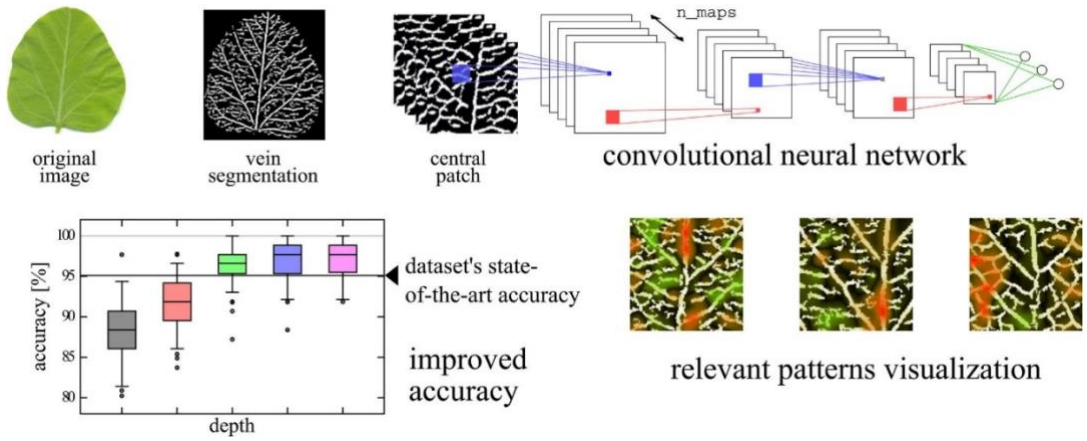
3.1.4. Termésminőség

A gépi tanulással támogatott MI-megoldásokat széleskörűen használják a termésminőség előrejelzésére is. A termény minőségi jellemzőinek pontos felismerése és osztályozása növeli a bevételt és csökkentheti a pazarlást. *Zhang et al.* (2017) egy új módszert mutattak be a betakarítás során a gyapotszálakba ágyazódott botanikai és nem botanikai idegen anyagok felismerésére és osztályozására. A vizsgálat célja a minőség javítása volt, miközben a szálak károsodását is minimalizálták. *Hu et al.* (2017) a Korla illatos körték azonosítására és megkülönböztetésére fejlesztettek ki módszert, ami különböző szempontok szerint két kategóriába sorolta a körtéket. A szétválogatáshoz itt is a gépi tanulást használták, a hiperspektrális reflexiós képalkotással készült képekkel. *Maione et al.* (2016) a rizsminták földrajzi eredetének meghatározására fejlesztettek ki egy új eljárást. A módszer a minták kémiai összetevőit vizsgálta. A felépített adatbázis alapján gépi tanulással pontosították az algoritmust. A fő cél a rizs földrajzi eredetének meghatározása volt Brazília két különböző éghajlati régiója, Goias és Rio Grande do Sul esetében. Az eredmények azt mutatták, hogy a minták osztályozása szempontjából a kadmium, a rubídium, a magnézium és a kálium a négy legfontosabb kémiai komponens.

3.1.5. Fajok felismerése

A mezőgazdasági növények automatikus felismerése és osztályozása nagy segítség a szakértők kiváltására, valamint az osztályozási idő csökkentésére. *Grinblat et al.* (2016) három hüvelyes növényfaj, nevezetesen a fehér bab, a vörös bab és a szójabab azonosítására és osztályozására szolgáló módszert mutattak be a levélerezet-minták

alapján. Az erezetmorfológia pontos információt hordoz a levél tulajdonságairól. A színnel és az alakkal összehasonlítva ideális eszköz a növények azonosítására. A rendszer vázlatát a *3.3. ábrán* láthatjuk.



3.3. ábra: Növényfajok azonosítása mélytanulás segítségével. Összefoglaló ábra (*Grinblat et al.,* 2016)

3.2. Gépi tanulás az állattenyésztésben

Az állattenyésztésen belül további két kategóriára oszthatjuk a kutatásokat. Az egyik főleg az állatjóléti megfigyelésekkel foglalkozik, míg a másik a klasszikus értelemben vett állattenyésztésről és a fejlesztésekről szól. Az állatjólét az állatok egészségével és jólétével foglalkozik, ezen a területen a gépi tanulás fő alkalmazási lehetősége az állatok viselkedésének megfigyelése a betegségek korai felismerése érdekében. Az állattenyésztési területen a legtöbb tanulmány a termelési rendszerrel kapcsolatos kérdésekkel foglalkozik. Az állatok különböző biológiai és fizikai jellemzőinek becslése, a tejhozam vagy éppen egyéb gazdasági jellemzők becslése és nyomon követése kiemelkedően fontos terület. A szarvasmarhák pontos súlyának meghatározása a legtöbb tenyésztő számára fontos szempont, az állomány fejlődésének monitorozására, illetve a takarmány összeállításánál és a vágási időpont eldöntésénél is. Az állatok rendszeres mérése hagyományos mérlegelési módszerekkel meglehetősen körülményes feladat, és az állatok számára is stresszes helyzetet teremt. Ehhez a feladathoz több szakember kell, és sok időt is vesz igénybe jelentős állatállomány esetében. Ezért az élősúly becslése fontos feladat, amit a digitalizáció és a mesterséges intelligencia együttműködésével, érdemes lenne a lehetőségekhez képest automatizálni.

Számos tanulmány foglalkozik az élősúly gépi becslésének eredményeivel. *Gemma et al*. (2019) tanulmányában arról számolt be, hogy a 3D képalkotási technika és a gépi tanulás segítségével jó eredményt értek el az élősúly becslésre (R^2 = 0.7, RMSE = 42 kg). Kísérletükben rögzített kamerarendszer segítségével 3D képet készítettek az állatokról. Ezekből a képekből 60 különböző méret- és területparamétert határoztak meg az állatokon, és ezen paraméterekből próbálták megbecsülni az állat súlyát, MI-algoritmusosok használatával.

3.2.1. Állatjóléti kutatások

Számos korábban megjelent tanulmány foglalkozik állatjóléti kutatások ismertetésével. *Dutta et al.* (2015) a szarvasmarhák viselkedésének ML-modelleken alapuló osztályozására mutatnak be egy módszert, amely mágneses érzékelőkkel és háromtengelyű gyorsulásmérővel ellátott nyakörvszenzorokkal gyűjtött adatokat használt. A vizsgálat célja különböző betegségek előrejelzése volt a szarvasmarhák táplálkozási szokásainak változása alapján.
A kísérlet során a mozgásminták ismeretében jó megközelítéssel meg tudták határozni, hogy az állat éppen mit csinál. A *3.1. táblázat* az egyes viselkedésminták adatalapú felismerésének mérőszámait mutatja.

3.1. táblázat: A szarvasmarhák viselkedésének felismerése Dutta et al. kísérletében

Behaviour class	Accuracy (%)	Specificity (%)	Sensitivity (%)	F1 score (%)	FDR (%)
Grazing	94	83	96	84	15
Searching	89	72	90	69	35
Ruminating	96	87	98	89	8
Resting	96	90	98	91	8
Scratching etc.	89	68	92	69	31
Average performance	93	80	95	80	19

Pegorini et al. (2015) a borjak rágási mintáinak automatikus azonosítására és osztályozására fejlesztettek ki egy rendszert. Kutatásukban egy gépi tanulás segítségével készült algoritmus a különböző takarmányok, a széna és a rizs rágási jeleinek adatait alkalmazta, kombinálva olyan viselkedési adatokkal, mint a kérődzés

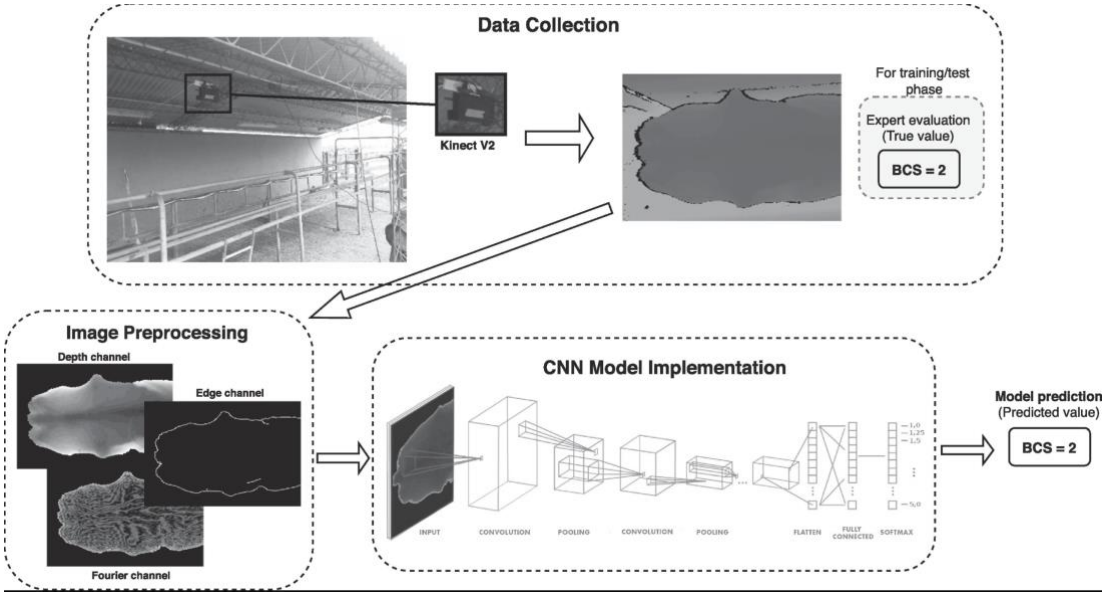
és a tétlenség. Az adatokat ebben az esetben optikai érzékelőkkel gyűjtötték. Kísérletükben a rágás 5 különböző fázistát 94%-os pontossággal tudták automatikusan elkülöníteni. *Matthews et al.* (2017) szintén gépi tanulás segítségével készítettek egy megfigyelő rendszert az állatok viselkedésének nyomon követésére. 3D mélységi megfigyelő kamerákat használtak a sertések mozgásának elemzéséhez. Az állatok különböző tevékenységeit (állás, mozgás, etetés és ivás) tudták elkülöníteni egymástól, és ezek alapján készítettek becsléseket a várható állatjóléti eseményekre.

3.2.2. Kondícióbírálati pontok gépi meghatározása

A testkondíció-pontozás módszerét használjuk a marhák faggyú-/zsírtartalékának és energia-egyensúlyának a jellemzésére. A rendszeres testkondíció-pontozás igen fontos a tenyésztés hatékonyságának felmérésére, az értéke befolyásolja a tejtermelést, a reprodukciót és az állat egészségét is. Ilyen módon kihatással van az állatok takarmányának összetételére, mennyiségére, végső soron pedig az egész gazdaság hatékony működésére. A kondíció elbírálásának módszerétől függetlenül fontos azt tudnunk, hogy mikor tarthatók a tehenek gyengébb minőségű takarmányon, mikor kell esetleg javítani a takarmány minőségét, beltartalmát, vagy mikor kell a meglevő szintet fenntartani. Ehhez pontosan meg kell határozni az állattermelés ciklusát és azt, hogy hogyan változtassuk a takarmányt a szaporodásbiológiai tulajdonságok javítása érdekében.

A testkondíció-pontozásra általában különböző mérések és vizsgálatok után, szakember becslése révén kerül sor. Ennek a szakértői pontozásnak a számítógéppel való segítése is ígéretes fejlesztési lehetőségnek számít.

Rodrigez et al. (2018) tanulmányukban bemutattak egy olyan eljárást, amelyben fixen rögzített kamerákkal készített képek felhasználásával próbálták teheneknél megbecsülni a testkondíció pontszámait. A rendszer vázlatát a *3.4. ábrán* láthatjuk. Az elkészített képeket egy konvolúciós neurális háló segítségével dolgozták fel és tanították meg a rendszert pontozni. Az elért eredmény 80%-os pontosságú volt.



3.4. ábra: A testkondíció-becslő rendszer vázlata
(*Rodrigez et al.*, 2018)

Később egy másik tanulmányban *Rodrigez et al*. (2019) ismertették a továbbfejlesztett változatot, ami az áttanítás (transfer learning – korábbi tanulási eredmények felhasználása), valamint az együttes tanulás (ensemble learning – több algoritmus együttes alkalmazása) módszerek segítségével tovább növelte a pontosságot, egészen 97%-ig.

Láthatjuk tehát, hogy a modern MI-módszerek megfelelően alkalmazva igen pontos becslést tudnak adni, korábban csak szakemberek által meghatározható, minőségi jellemzőkre is.

3.2.3. Tejminták elemzése

Az elmúlt években rohamosan nőtt a tesztelés vagy már bevezetés fázisában lévő MI-rendszerekkel is támogatott okosfarmok száma. A legtöbb esetben különböző érzékelőket szerelnek az állatokra. Ezek az érzékelők mérik a mozgást és a testhőmérsékletet, valamint akár távolról is segítik az egyes egyedek felismerését és ezzel pontos, egyedre lebontott adatok gyűjtését. A hőmérséklet mérésévével gyorsan lehet következtetni az esetleges betegségek kialakulására. A nagy mennyiségű adatot szintén MI segítségével dolgozzák fel, hogy a rendellenes mintázatok felismerésével megfelelő következtetéseket tudjanak levonni a jövőre nézve.

A precíziós állattenyésztés elengedhetetlen feltétele az állatok pontos egyedi azonosítása. *Yongliang Qiao et al.* (2019) bemutattak egy olyan rendszert, ami egyetlen kamera képét felhasználva sikeresen azonosította egy több száz marhát számláló tenyészet egyedeit a mesterséges intelligencia segítségével.
Az *Advanced Animal Diagnostic* cég többek között tejminták elemzésével foglalkozik. Egyik hordozható berendezésük képalkotó eljárással vizsgálja a fehérvérsejteket vérből vagy tejmintából *(3.5. ábra)*. A rendszer megállapítja bizonyos típusú fehérvérsejtek mennyiségét, majd a mélytanulás módszerének alkalmazásával az adatok alapján egészségügyi következtetéseket képes levonni.

3.5. ábra: A Qsout Farm Lab hordozható diagnosztikus berendezés, amely egy beépített mikroszkóppal vizsgálja a tejmintákat és laboratóriumi minőségű adatokat képes szolgáltatni helyben (a kép forrása: Advanced Animal Diagnostic)

Hasonló elven működik a *SomaDetect* nevű startup tejanalizáló módszere is. Optikai szenzorok segítségével képet állítanak elő a tej fényszórásának rögzítésére *(3.6. ábra)*. Ezeket a képeket aztán egy konvolúciós neurális hálón futó mélytanuló algoritmus segítségével dolgozzák fel. Korábban laboratóriumi körülmények között állítottak elő nagy mennyiségű adatot, és azt a modell tanításához használtak fel. Az így rendelkezésre álló adatok alapján akár a helyszínen részletes egészségügyi vagy éppen tejminőségi becsléseket tudnak készíteni, amelynek jelentősége felbecsülhetetlen gyakorlati szempontból.

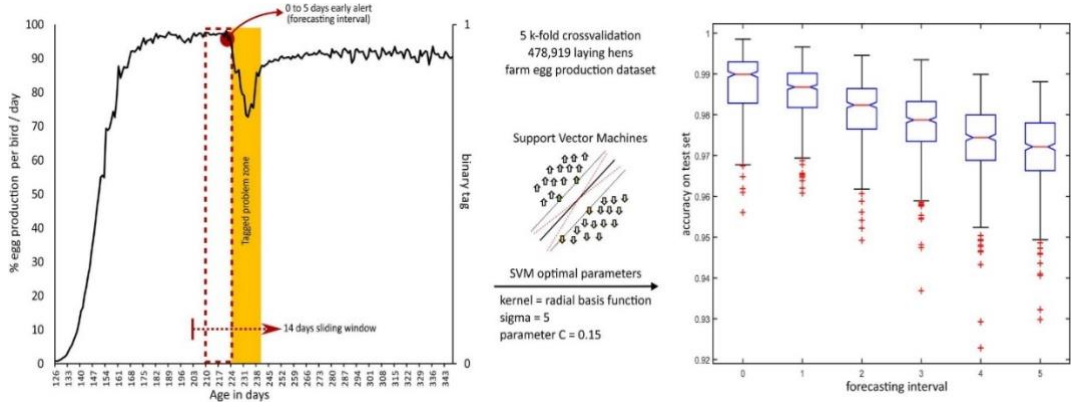
3.6. ábra: A SomaDetect cég érzékelője tejfeldolgozó rendszerbe építve
(a kép forrása: SomaDetect)

Az állatok viselkedésének megfigyelésére és elemzésére is egyre nagyobb mértékben használják az IoT-eszközöket, ahol az adatokat aztán megfelelő MI-algoritmus segítségével dolgozzák fel. Az állatok viselkedésének megfigyelésével és a kellő időben történő beavatkozással sokkal könnyebb biztosítani a stresszmentes környezetet, és ezzel az egész tenyészet kezelése egyszerűbb lesz, és bizonyítottan nő a termelékenység is.

3.2.4. Állattenyésztési előrejelzések

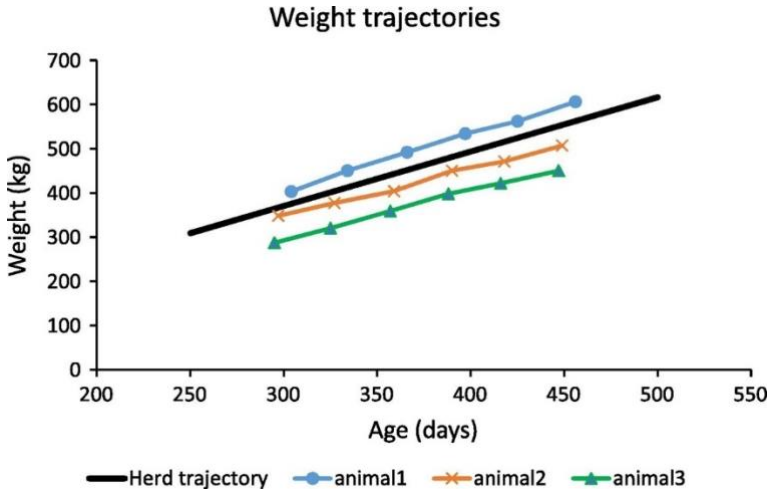
Az állattenyésztési rendszerek kialakításánál és tervezésénél is számos esetben használják már a gépi tanulásra alapozott algoritmusokat. Általában a gazdaság hatékonyságának optimalizálása érdekében a gazdálkodási paraméterek pontos előrejelzésére és becslésére készültek kutatások. *Craninx et al.* (2008) munkájukban a bendőfermentációs mintázat tejzsírsavakból történő előrejelzésére alkalmas módszert mutattak be. A neurális hálót használó algoritmus fő célja a bendőfermentáció minél pontosabb előrejelzése volt, mert ez jelentős szerepet játszik a tejtermelés előrejelzésében. A kutatás azt is bebizonyította, hogy a tejzsírsavak ideális tulajdonságokkal rendelkeznek a zsírsavak bendőben lévő moláris arányainak előrejelzésére. *Morales et al.* (2016) vizsgálata a tyúktartáshoz kapcsolódott. Ők egy SVM-modellen (Support Vector Model) alapuló módszert mutattak be a kereskedelmi tojástermelésben jelentkező hibák korai felismerésére és előrejelzésére. Az általuk kifejlesztett algoritmus kielégítő pontossággal 0–3 nappal előre jelezte a termelési görbe változásait. Általában azt tapasztalták, hogy 1 nappal korábban 90%-nál

nagyobb pontossággal tudták előre jelezni a változásokat. A kutatás összefoglalóját a *3.7. ábrán* láthatjuk.



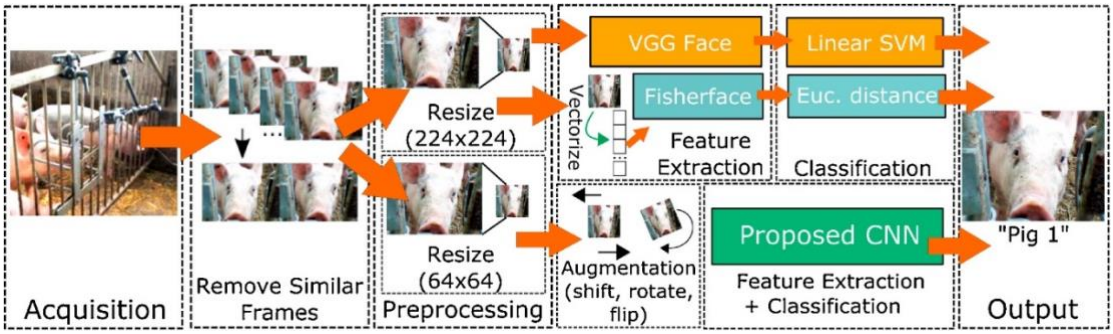
3.7. ábra: Tojástermelésben jelentkező hibákat felismerő rendszer elvi működési modellje (*Morales et al.,* 2016)

Hasonló módszert (SVM) használtak *Alonso et al.* (2015) a szarvasmarhák súlygyarapodásának pontos időbeli becslésére *(3.8. ábra)*. A szarvasmarhák súlyának pontos becslése nagyon fontos a tenyésztők számára a megfelelő gazdasági számítások elvégzéséhez. Cikkükben az Asturiana de los Valles fajtájú húsmarhák testtömegének előrejelzésére szolgáló függvény kifejlesztésével foglalkoztak. Meghatározták a teljes tenyészetre vonatkozó függvényt és ehhez tudták hasonlítani az egyes egyedek súlyának változását néhány mérés alapján. Az eredmények azt mutatják, hogy a bemutatott módszer képes a vágás előtt 150 nappal a test súlyának előrejelzésére.



3.8. ábra: *Alonso et al.* súlygyarapodást előre jelző függvényeinek meghatározása az *Asturiana de los Valles* szarvasmarhafajtában

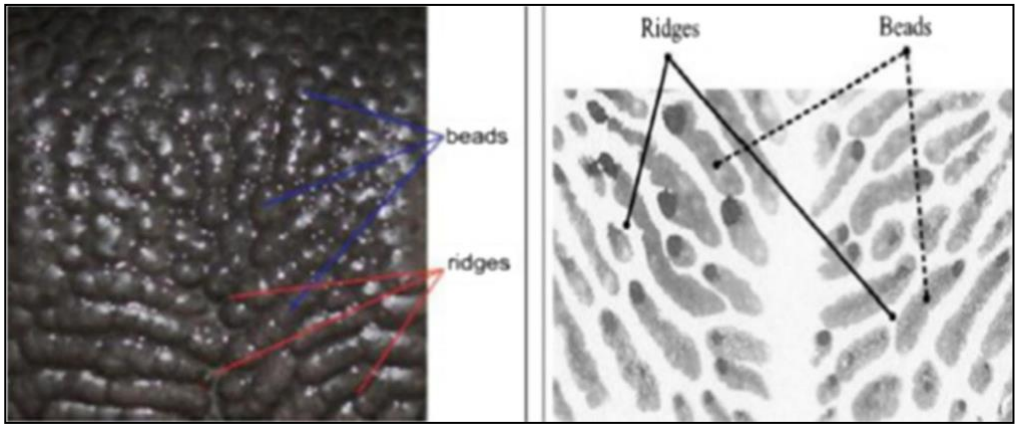
Hansen et al. (2018) egy digitális képeken alkalmazott, konvolúciós neurális hálózatokon (CNN) alapuló módszert mutattak be sertések felismerésére. A kutatás fő célja az állatok azonosítása volt rádiófrekvenciás azonosító (RFID) címkék nélkül, amelynek viselése az állat számára kellemetlen lehet, hatótávolságuk korlátozott, a módszer pedig időigényes. A képek elkészítéséhez webkamerát használtak és az így gyűjtött adatokkal egy CNN-alapú modellt tanítottak fel. A rendszer vázlata a *3.9. ábrán* látható. Az így elkészült algoritmust élőben is kipróbálták, és 96,7%-os pontossággal ismerte fel a sertéseket.



3.9. ábra: A *Hansen et al.* által fejlesztett CNN-alapú rendszer működésének vázlata

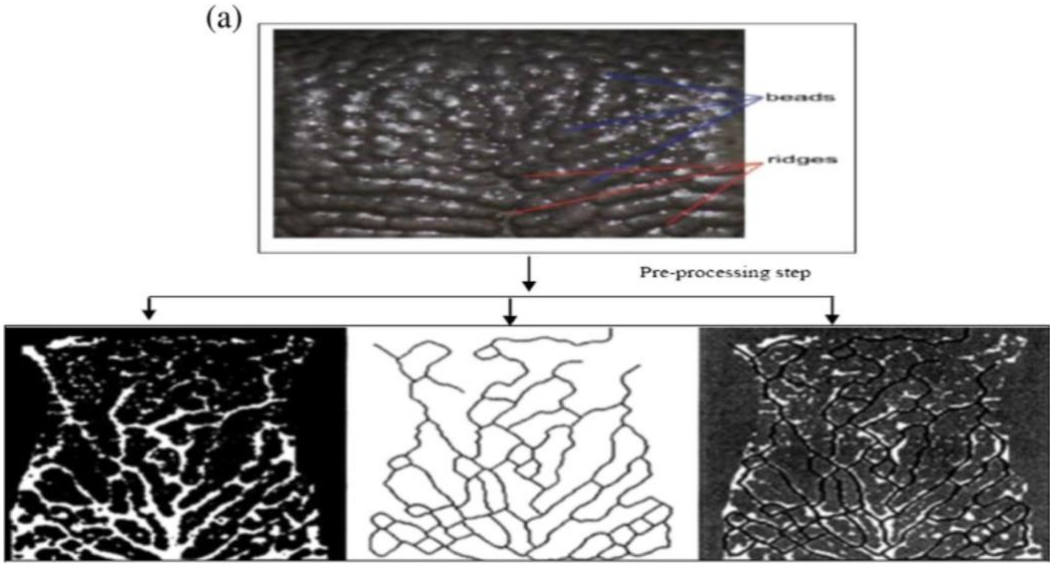
3.2.5. Szarvasmarhák azonosítása orrnyílásaik alapján

Az állatbiometria egy feltörekvő kutatási terület, amely az állatok vizuális megjelenésének az általános jellemzők és elsődleges biometrikus jellemzők alapján történő ábrázolására épül. Az egyes szarvasmarhák azonosítása világszerte fontos kérdés a különböző fajták osztályozása, nyilvántartása, nyomon követhetősége, egészségügyi menedzsmentje szempontjából (*Kuhl HS, Burghardt T*, 2013). A klasszikus állatfelismerési módszereknek komoly problémái vannak, mivel ezek mind kézi azonosítási rendszerek. A klasszikus állatazonosítási módszerek könnyen hamisíthatók a felcímkézett füljelzőkön található egyedi számok megváltoztatásával. Megoldás lehet a szarvasmarhák orráról készített pontszerű képek használata. Az orrpontalapú azonosítás hasonló az emberi ujjlenyomatok pontjainak felismeréséhez. A kutatásban 500 szarvasmarhát vizsgáltak, egyenként 10 darab 20 megapixeles képeket készítve róluk. A képeken ezután két fő karakterisztikus bélyeg (beads, ridges) alapján azonosították az állatokat *(3.10. ábra)*.



3.10. ábra: A két fő karakterisztikum
(*Kumar et al.,* 2017)

A képeket ezután feljavították úgy, hogy mérsékelték a zajokat, javítottak a képminőségen és láthatóbbá tették a fő jellemzőket. Az így kapott modelleket pedig a mesterséges intelligencia fel tudta ismerni, és dolgozni tudott velük *(3.11. ábra)*.

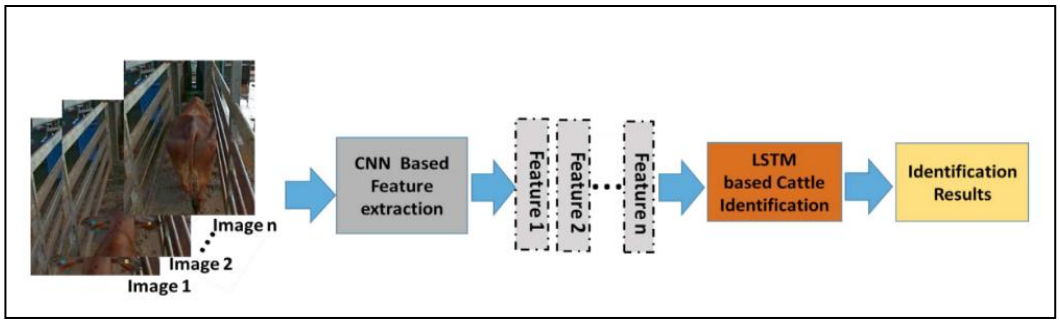


3.11. ábra: Feldolgozás utáni modellek
(*Kumar et al.,* 2017)

A kutatás során 5000 képpel dolgoztak, amelynek a végén 96,74%-os pontossággal a program azonosítani tudta a szarvasmarhákat. Úgy gondolom, hogy a nagy pontosság ennek az eljárásnak a megalapozottságát igazolja.

3.2.6. Szarvasmarhák azonosítása küllem alapján

Korábban már bemutattam egy küllem alapján működő azonosítási megoldást. Mivel ez igen fontos feladat, ezért több megoldás is létezik. Egy precíziós gazdálkodás megfelelő működtetéséhez szükség van olyan modern megoldásokra, mint a szarvasmarháink egyedi azonosítása. Ez jelenleg a leggyakrabban vizuálisan, egyedi rádiófrekvenciával vagy füljelzőkkel történik. Ezeknek a rendszereknek a helyettesítésére pedig újból segítségül lehet hívni a mesterséges intelligenciát. Egy kutatásban pontosan erre tettek kísérletet olyan mélytanulási rendszer létrehozásával, amely CNN *(Convolutional Neural Network)* és LSTM *(Long Short-Term Memory)* hálózatok kombinálásával képes azonosítani az állatokat. A CNN és az LSTM hálózatra azért volt szükség, mivel hatékonyan tudnak vizuális forrásból térbeli információkat kinyerni, majd ezeket modellezni. Összesen 516 videófelvétel készült 41 szarvasmarháról három hónapon keresztül, havonta egyszer. Először a rögzített felvételeket képkockánként a CNN hálózat kielemezte, majd minden egyes szarvasmarháról készített egy olyan halmazt, amelyben az állatok vizuális jellemzői, illetve az egyedi mozgása szerepel. Ezután a kapott CNN halmazokat beviszik az LSTM hálózatba, amely képes megtanulni, majd újból felismerni a szarvasmarhákat *(3.12. ábra)*.



3.12. ábra: Feldolgozás utáni modellek
(*Yongliang Qiao et al.,* 2019)

A kísérleti eredmények azt mutatják, hogy a módszer 88%-os és 91%-os pontosságot ért el 15 és 20 képkockás videók esetén, tehát a kísérlet mindenképpen sikeresnek mondható. Emellett pedig, ha LSTM hálózat nélkül, csak a CNN rendszerrel dolgoztak, akkor a közel 90%-os pontosság helyett csak 57% lett az eredmény. Ez is mutatja, hogy több, mesterséges intelligenciára épülő modell összekapcsolása javítja az elérhető pontosságot.

3.2.7. Szarvasmarhák UAV-képek alapján történő észlelése

A pilóta nélküli légi járművek egyre jobban használható eszközökké válnak a gazdaságokban. Ez a fajta technológia különösen hasznosnak bizonyul az extenzív szarvasmarhatartás tekintetében, mivel a termelési területek általában kiterjedtek, és az állatokat lazábban felügyelik. A drónok folyamatos fejlődésével és a konvolúciós neurális hálózatok elterjedésével egyre hatékonyabban lehet releváns információkat kinyerni digitális képekből. Egy 2019-es brazil tanulmányban azt vizsgálták, hogy mennyire pontosan lehet UAV által készített digitális képek alapján felismerni az egyes állatokat. Itt a felismerésen volt hangsúly, mivel, ha beazonosították az egyedet, utána különböző feltételezések vonhatók le, mint például az állatlétszám, rendellenes események vagy a testméretek felvétele. A tanulmány során 1853 képet vizsgáltak meg, amelyek Chanchim szarvasmarhákról, felülnézetből készültek *(3.13. ábra)*.

3.13. ábra: Az állatokról készült UAV-felvételek a Chanchim fajtában (*Barbedo et al.,* 2019)

Az állatokról nem csak hibátlan minőségű képek készültek. Összesen 15 programot teszteltek, hogy különféle fényviszonyok, időjárási feltételek, túl magas fényerő és elmosódás mellett milyen eredményességgel működnek. Átlagos 95%-os pontossággal dolgoztak, mégis a NasNet Large volt a legeredményesebb 99,2%-kal. Az egyes modellek eredményei a *3.2. táblázatban* láthatók.

3.2. táblázat: A programok eredményei (*Barbedo et al.,* 2020)

CNN	Accuracy	Precision	Recall	F1 Score
VGG-16	0.972	0.973	0.973	0.970
VGG-19	0.973	0.973	0.973	0.975
ResNet-50 v2	0.977	0.978	0.978	0.975
ResNet-101 v2	0.983	0.985	0.985	0.985
ResNet-152 v2	0.967	0.970	0.970	0.965
MobileNet	0.983	0.980	0.980	0.983
MobileNet v2	0.787	0.855	0.790	0.778
DenseNet 121	0.852	0.895	0.868	0.865
DenseNet 169	0.935	0.943	0.933	0.935
DenseNet 201	0.935	0.945	0.938	0.938
Xception	0.969	0.968	0.968	0.968
Inception v3	0.979	0.975	0.975	0.975
Inception ResNet v2	0.983	0.983	0.983	0.985
NASNet Mobile	0.857	0.890	0.858	0.853
NASNet Large	0.992	0.993	0.993	0.995

3.2.8. A Pantaneira szarvasmarha felismerése CNN hálózattal

Az állatok azonosításának olcsó, gyors és biztonságos módszerei rendkívül fontosak a gazdálkodási rendszer számos feladata szempontjából. Ezek a rendszerek megkönnyítik az állattenyésztési folyamatokat, a tevékenységek ellenőrzését, és biztosítják a termelés számonkérhetőségét. A brazil állatállomány legelterjedtebb azonosítási rendszerei közé tartozik a forró vasalással, elektromos vagy tűztetoválással történő megbélyegzés, a füljelzővel és a nyakörvvel történő jelölés. Az azonosíthatóság érdekében egy brazil tanulmányban 2020-ban azt vizsgálták, hogy 4 kihelyezett kamera és az általuk készített képek alapján egy CNN rendszer milyen pontosan képes felismerni az adott állatokat. A rögzített képek a *3.14. ábrán* láthatók. Ehhez 51 állatot vizsgáltak a NUBOPAN központban, ahol hátulról, mindkét oldalról és szemből készítettek képeket, összesen 27 849-et.

3.14. ábra: A *Pantaneira* szarvasmarhákról készített fotók (*Weber et al.,* 2020)

Három különböző CNN rendszert hasonlítottak össze, melyek – amint a *3.3. táblázatból* is látszik – szinte 100%-os pontossággal dolgoztak.

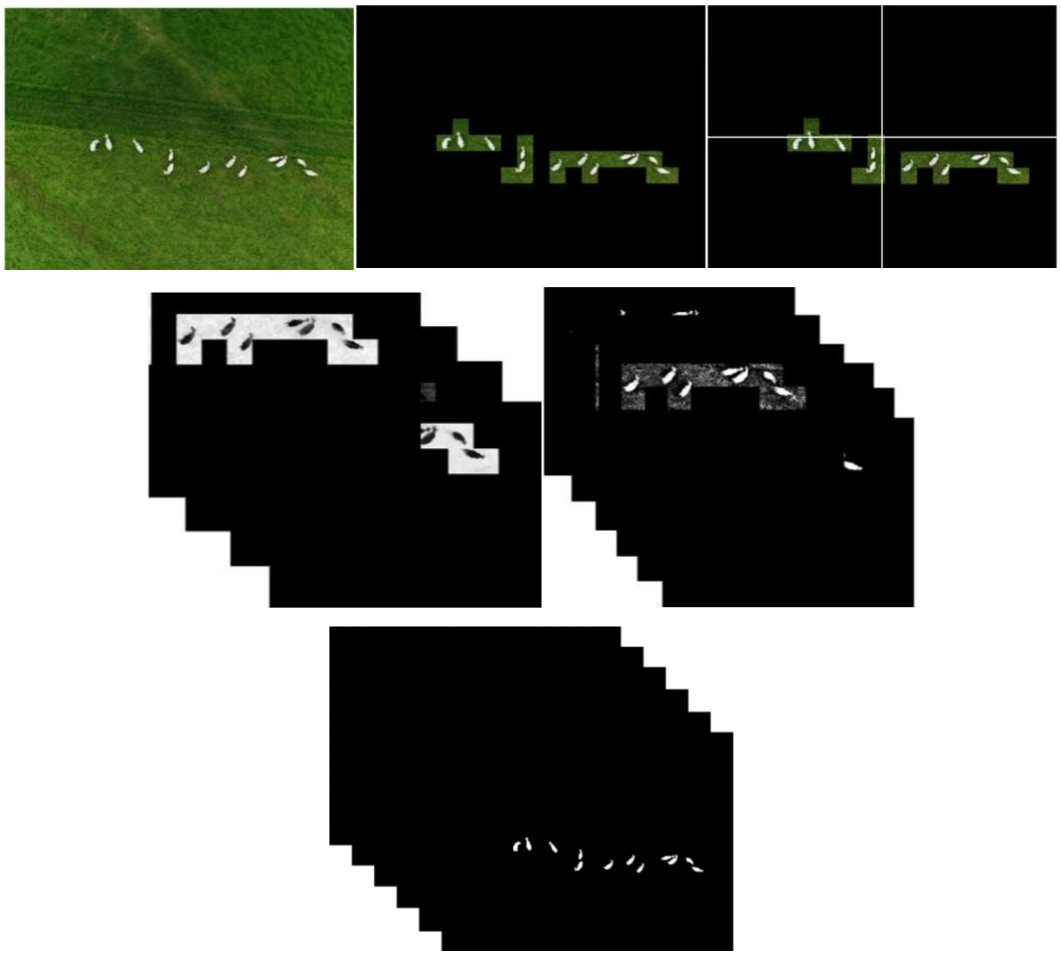
3.3. táblázat: A három rendszer teljesítménye %-ban (*Weber et al.,* 2020)

Model	Accuracy Training (%)	Accuracy Test (%)
ResNet50	99.20	99.78
InceptionResNetV2	98.87	99.52
DenseNet201	99.74	99.85

3.2.9. Nelore és Canchim szarvasmarhák számlálása UAV-képeken

A pilóta nélküli légi járművek vagy drónok elterjedésével nagy felbontású, alacsony költségű, madártávlatú képet lehet készíteni a birtokról. Bármennyire is ígéretes ez a technológia, a benne rejlő lehetőségek teljes körű kiaknázását akadályozza, hogy a releváns információk kinyerése az UAV segítségével rögzített képekből korántsem egyszerű. Az állatállomány becslése esetében nehézséget okoz az állatok mozgása, a táj változatossága (csupasz talaj, száraz legelő), az akadályok, például fák és istállók által okozott holttér, valamint az állatok csoportosulása. Egy 2020-as brazil tanulmány próbálkozást tett, ezeknek a problémáknak a megoldására. Az első lépésben minden képet szabályos rácsháló segítségével négyzetekre osztottak, és az állatok részeit

tartalmazó négyzeteket egy CNN segítségével azonosították *(3.15. ábra)*. A második lépésben színtér-manipulációkat alkalmaztak az állatok és a háttér közötti kontraszt növelésére, és küszöbértékeket állítottak be bináris maszkok létrehozására. A harmadik lépésben a maszkokat bináris műveletekkel kombinálták, és morfológiai műveleteket alkalmaztak a klaszterezett állatok elkülönítésére. A negyedik lépésben a képeket összevetették a korábban osztályozott képekkel, és így megkapták a végső becsléshez szükséges információkat.



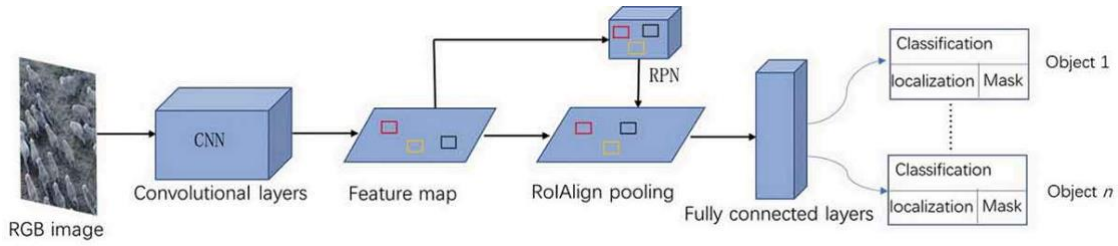
3.15. ábra: A Nelore és Canchim szarvasmarhákról készült UAV-felvételek feldolgozása (*Barbedo et al.,* 2020)

Az osztályozásért felelős betanított CNN-modellnek a NasNet Large modellt választották, amely 98% feletti pontosságú meghatározást biztosított az állatok felismerésében.

3.2.10. Állatok osztályozása és számlálása drónfelvételeken Mask-R-CNN rendszer alapon

A drónokkal történő állományelemzés egy másik módszerét mutatták be kínai és mongol szerzők egy 2020-ban megjelent cikkükben (*Beibei et al.,* 2020). Ez a kutatás olyan költséghatékony megoldást mutat be az állomány számlálására és osztályozásra, amely segíti különböző mérések és állatjóléti megoldások kidolgozását. A dolgozatban alkalmazott módszer a legkorszerűbb mélytanulási technikát, az úgynevezett Mask R-CNN-t alkalmazza a tulajdonságok kinyeréséhez és a tanuláshoz.

A Mask R-CNN alapvetően egy konvolúciós neurális háló, aminek az erőssége a képszegmentálás *(3.16. ábra)*. A mélytanuló neurális hálók olyan változatáról van szó, amely jó minőségű maszkot készít a kép minden részeltéhez és ezzel megkönnyíti az egyedek megszámlálását a képen.



3.16. ábra: A kutatásban használt Mask-R-CNN rendszer vázlata (*Beibei et al.,* 2020)

A kísérlet során drónokkal készítettek videófelvételt a mezőn legelő nyájról. Ezeket utána azonos 510 × 512 pixel méretű állóképekké vágták és ezt dolgozták fel a neurálisháló-alapú rendszerrel.

A kísérletben használt rendszer értékelésére valós (hagyományos) számláláson alapuló rendszereket használtak, és a kísérleti eredmények azt mutatják, hogy a drónokkal készült képek kiértékelése 96%-os pontossággal képes osztályozni az állatállományt, a juhok és szarvasmarhák számát pedig 92%-os pontossággal határozta meg a rendszer.

3.2.11. MI által támogatott testkondíció-értékelés

A testkondíció-értékelési (BCS) rendszer a szarvasmarhák kategorizálására alkalmas, valamint felhasználják az árképzéshez, illetve a jövőbeni hasznosíthatóság megállapításához is. Szoros kapcsolat áll fenn a borjú születésekor kapott testkondíció-

érték, és az azt követő első 90 nap után kapott értékelés eredménye, a borjú egészsége és jövőbeni potenciálja között.

Korábbi tanulmányok kimutatták, hogy a fenotípus és a várható utódok kapcsolata erősebb a húshasznú marhák esetében, mint más tulajdonságoknál (*Tőzsér et al.,* 2020). A vizuálisan értékelt húsmarha-tulajdonságok különféle számban öröklődnek. Korábbi kutatásokból tudjuk azt is, hogy mivel szoros korreláció áll fenn (> 0,70) az izomzat és a vágási tulajdonságok között, biztonsággal alkalmazhatunk különböző osztályozási módszereket a gyakorlatban is (*Journaux et al.,* 1994; *Korchma et al.* 1986).

A testkondíció-pontozást széles körben alkalmazzák a szarvasmarha-tenyésztés egyéb területein is. Az új technikák, amelyek előre jelzik a testkondíciót, nagyon fontosak a gazdálkodók számára. Például egy kísérlet során BCS kameraképeket használtak a tejhasznú tehenek testkondíciójának és termelési paramétereinek kiszámítására (*Chlebowski et al.,* 2020).

A testkondíció-pontozást a szarvasmarhák zsírtartalékának és energiaháztartásának jellemzésére használják. A rendszeres testkondíció-pontozás a tenyésztés tervezhetőségét teszi biztonságosabbá, mivel következtetni lehet belőle a tejtermelés mértékére, a szaporodási képességekre és az állatok egészségére is. Ezek ismeretében pontosabban meg lehet határozni a szükséges takarmány minőségét és mennyiségét, akárcsak az egész gazdaság hatékonyságát. Fontos tudni, mikor tarthatók az állatok alacsonyabb minőségű takarmányon, és mikor szükséges javítani a takarmány minőségén és összetételén.

A testkondíció-pontozást általában különféle mérések és szakértői vizsgálatok alapján végzik. A szakértői pontozás eredményei alapján fejlesztett MI-modell hasznos kiegészítője lehet a pontozási eljárásnak.

VanderWaal et al. (2017) tanulmányukban bemutattak egy módszert, amely megkísérelte becsülni a tehenek testkondíció-pontszámát rögzített kamerákkal készített képek alapján. A rögzített képeket egy konvolúciós neurális hálózat segítségével dolgozták fel, és az így tanított rendszerrel elért eredmény körülbelül 80%-os pontosságú volt.

Ez is bizonyítja, hogy a modern MI-módszerek ötletes alkalmazása még az olyan minőségi jellemzők esetében is nagyon pontos becslést adhat, amelyeket korábban csak szakemberek tudtak meghatározni.

3.3. Adatelemzés a szarvasmarha-tenyésztésben

Természetesen nem csak képfeldolgozással és szenzorok segítségével jelenik meg a mesterséges intelligencia a szarvasmarha-tenyésztésben. Az adatelemzés tudományának eredménye is megjelennek a precíziós tenyésztési technológiák eszköztárában. A genetikai és fenotípus-megfigyelések régóta folynak hagyományos statisztikai módszerekkel. Azonban a mesterséges intelligencián alapuló adatelemző módszerekkel a jövőre adható becslések pontossága és sebessége is fejlődött.

A *Teheráni Egyetem* (Irán) és a *Wisconsini Egyetem* (USA) szerzőinek közös kutatása a tejelő szarvasmarhák tenyészértékének mesterséges neurális hálózat és fuzzy rendszer segítségével történő becslésének lehetőségét vizsgálja. Tanulmányuk célja, hogy ezeknek a tanulási módszereknek a lehetőségeit mérjék fel annak érdekében, hogy az iráni tejelőszarvasmarha-állományra vonatkozó becsült tenyészértékeket kapjanak (*Shahinfar et al.,* 2012). A tanulmányban felhasznált adatokat az *Iráni Állattenyésztési Központ* (ABCI, Teherán) gyűjtötte, az adatbázis 119 899 elemet tartalmazott az 1990 és 2005 között időszakból, a 22 és 36 hónapos kor között ellett Holstein tehenek laktációjáról. A kutatáshoz felhasznált adatok a tej- és zsírhozamra, valamint a környezeti feltételekre, például a hőmérsékletre, a páratartalomra vonatkoznak.

A mesterséges neurális hálózatok *(ANN)* adatalapú tanulási modellek, ahol a modell a tanítási fázisban a megadott bemenetek és a megfelelő kimenetek alapján tanulja meg az adott modellarchitektúra paramétereit. Ebben az eljárásban az adatok a neurális hálózat összes rétegén keresztül terjednek, miközben a cél a kimeneti értékek elérése. A tanulási fázis során a háló figyeli a hibát a kimeneti értékek elérésében az adott bemenetekre vonatkozóan, és beállítja a paramétereket a bemeneti és kimeneti értékek közötti összetett kapcsolatok modellezése érdekében. A mesterséges neurális hálók használata nagyon hatékony lehet, és gyakran fedez fel ismeretlen kapcsolatokat a bemenetek és kimenetek között, mind az osztályozási, mind a regressziós problémákban. Az ANN-nek ez a mintafelismerő képessége tovább fokozható különböző technikákkal, például az adatok előválogatásával vagy előfeldolgozásával, az optimális hálózati architektúra és a képzési ciklusok optimális számának megtalálásával, a bemeneti jellemzők számának és kombinációjának változtatásával és a tanulási paraméterek értékeinek beállításával.

Shahinfar et al. (2012) egy feed forward backpropagation többrétegű perceptront használnak 1 bemeneti, 2 rejtett és 1 kimeneti réteggel. A bemeneti réteg neuronjai

megfelelnek az egyes magyarázó változóknak. A rejtett rétegekben lévő neuronok tangens hiperbolikus (tanh) aktivációs függvényt használnak a bemeneteik súlyozott összegének kialakítására. A kimeneti neuronok ugyanezt az aktivációs függvényt használják a második rejtett réteg kimeneteinek súlyozott összegének kialakítására és a végső kimeneti értékek kiszámítására. A tanh függvény értékkészlete a (−1, 1) intervallum, és a függvény differenciálható, ami jó a backpropagation algoritmus számára. A backpropagation (visszaterjesztés) lényege, hogy kiszámítja, hogyan változzon minden súly a neurális hálózatban, hogy csökkentse a hibát. Ezt a gradiens segítségével teszi, amihez a függvény deriváltjára van szükség. Ez a függvény megfelelő a tenyészértékek becslésére is, amelyek mindig egy bizonyos érték körüli „±" intervallumban vannak.

Gota et al. (2018) tanulmányukban általánosságban vizsgálták meg a Big Data elemzések használhatóságát a precíziós állattenyésztésben. A képfeldolgozáson alapuló következtetéseken kívül még mindig a hatékonyan felhasználható adathalmazok meghatározásán van a hangsúly, ezért is fontos, hogy általánosságban is kifejezzék az ilyen típusú elemzések alkalmazhatóságát.

Tanulmányukban megállapítják: „*A precíziós állattenyésztést lehetővé tevő, teljesen automatizált adatgyűjtési vagy fenotipizálási platformot nemcsak a növekvő adatmennyiség, hanem a valós idejű adatgyűjtés összetett és dinamikus jellege is jellemzi. Az adatintenzív technológiák támogatásával folyamatosan nyomon követhetjük az állatokat a tenyésztési folyamat során, és ezeket az információkat felhasználhatjuk az egészség, a jólét, a teljesítmény és a környezeti terhelés javítására."* Az állattenyésztő közösség ma gyakran nincs annak az infrastruktúrának és azoknak az eszközöknek a birtokában, amelyekkel az új típusú adatokat teljes mértékben ki tudják használni. Molekuláris szintű információkkal, például a genomikával vagy transzkriptomikával kombinálva az új gépi tanulási és adatbányászati technikák előremozdíthatják a precíziós állattenyésztés megvalósítását, hogy a nagy mennyiségű adatból kritikus információkat nyerjenek ki és előre jelezzék a jövőbeli megfigyeléseket. Az adatok befogadására, biztosítására és megosztására szolgáló kiberinfrastruktúra szintén felhasználható a Big Data kiaknázására. Várhatóan a prediktív Big Data egyre inkább elterjed majd valamennyi állattenyésztési tudományágban. Azt állítjuk, hogy az első lépések ezen az úton az ilyen eszközök előnyeinek és hátrányainak beazonosítása, amikor azokat az állattudomány-specifikus területeken alkalmazzák. Ezen túlmenően az egymást kiegészítő háttérrel rendelkező transzdiszciplináris területek, például az informatika, a közgazdaságtan, a mérnöki

tudományok, a matematika és a statisztika, valamint az ipar szoros együttműködése elengedhetetlen a nagy áteresztőképességű és heterogén adatok elemzésére szolgáló élvonalbeli megközelítések hatékony kifejlesztéséhez. A precíziós állattenyésztés lehetővé teszi a gazdák számára, hogy gyorsan képesek legyenek beavatkozni a tenyésztési folyamatokba, valamint a nagy mennyiségű adattal dolgozó mezőgazdaság mesterséges intelligencián alapuló fejlesztése nagy segítség lehet az állattudományok előtt álló kihívások kezelésében is.

3.4. Hazai MI-alkalmazási területek

Hazai környezetben is megkezdődött a tenyészetek digitalizációja, és így egyre szélesebb körű lehetőség nyílik az MI-technikák bevezetésére. Egyre több adatot tudunk gyűjteni nemcsak kézi mérési módszerekkel, hanem az IoT-eszközök terjedésével emberi beavatkozás nélkül is.

A fejlesztéseket pályázatokkal is ösztönzik a különböző felelős szervek. 2020-ban a NAK *TechLab* Egyetemi Ötletversenyen egy olyan MI-fejlesztés lett a nyertes, amely fejlett képfeldolgozási módszerekkel segítheti a szarvasmarha-sántaság elleni fellépést.

A képfeldolgozás és okosérzékelők mellett azonban érdemes lenne a nagy mennyiségű egyéb adat feldolgozásában is megkeresni az MI-algoritmusok helyét, hiszen a tenyészértékbecslés a hazai termelők számára is létfontosságú feladat. Jelenleg ezek a becslések komoly szakértői munkát igényelnek, és ennek a munkának az informatikai segítése – ahogy azt már *Saleh Shahinfar et al.* (2012) jelezték – a kutatók következő célterülete. A szarvasmarhák típusának értékelése is számos adat és paraméter alapján lehetséges, itt ismegkerülhetetlen a fejlett MI-algoritmusok szerepe.

Jó példa további felhasználási területre a tejtermelő tehenek különböző ketózis típusba sorolásának támogatása MI-módszerekkel. Még jelenleg is a szakemberek különböző labor- és vizuális adatok alapján próbálják besorolni a teheneket a megfelelő ketózis típusba. A megfelelő besorolás nagyon fontos a helyes beavatkozás megtervezéséhez (*Monostori és Dégen,* 2019).

3.5. A szakirodalom kritikai elemzése

A szakirodalom áttekintése alapján elmondható, hogy a mesterséges intelligencia használatának fontossága egyértelműen látszik a mezőgazdasággal foglalkozó kutatók számára. A cikkek túlnyomó többsége képi adatok elemzésével foglalkozik, és ebben

látható a legnagyobb fejlődés is. Ez egyrészt érthető, hiszen ezek a rendszerek képesek a leginkább kiváltani a terepen dolgozó szakértők szerepét. Ahol sokszor a tapasztalt tenyésztők szakértelme vagy éppen nagyszámú ember fárasztó munkája javította a termelés hatékonyságát. Az emberi munkaerő kiváltása intelligens rendszerekkel jelentős költségmegtakarítás a gazdaságok számára, ezért is népszerűek az ilyen jellegű kutatások. Ugyanakkor fontos hangsúlyozni, hogy a meglévő adatbázisok elemzése, illetve az egyéb szenzorok által szolgáltatott adatok begyűjtése és feldolgozása is nagy érték a gazdaságok számára. Az is egyértelművé vált számomra, hogy a fellelhető szakirodalom (és a tudományos kutatások) az adatelemzésen alapuló predikciós algoritmusok és eljárások vonatkozásában még számos területen korlátozott eredményeket hoztak, a jövőbeni fejlesztések itt meghatározóak lesznek.

4. ANYAG ÉS MÓDSZER

Kutatásaim során – a célkitűzésekkel összhangban – három különböző becslési, predikciós modell feltanítását, ellenőrzését és validálását végeztem el.

1. Szarvasmarhák küllemi bírálati pontszámainak becslése
2. Szomatikus sejtszám becslése tejmintákban
3. Kazein mennyiségének becslése tejmintákban

A hipotézisem az volt, hogy meglévő adatbázisok segítségével feltanított, gépitanuláson alapuló modellekkel az említett becslések jól elvégezhetők. Fejleszthető tehát olyan automatikus MI-rendszer, amely folyamatosan tanulva képes egyre jobb minőségű becsléseket adni a szarvasmarha-tenyésztők számára fontos területeken.

A kutatás első fázisában az elérhető adatbázisok felkutatása volt a feladat. Meg kellett keresni azokat a hitelesített és létező adatbázisokat, amelyek használhatók egy jó minőségű modell elkészítéséhez.
Számos tenyésztőegyesület *(Holstein-fríz Tenyésztők Egyesülete, Magyartarka Tenyésztők Egyesülete)* végez szakértők segítségével szarvasmarha küllemi bírálati pontozást. Ezek a pontozások általában egységes módszer alapján, arra speciálisan képzett szakértők segítségével történnek. Egy törzstenyészet állatait ennek segítségével tudják jól szelektálni *(4.1. ábra)*. A modellem tanításához a *Limousin és Blonde d'Aquitaine Tenyésztők Egyesületétől* (Budapest) kapott 325 állat pontozási eredményét tartalmazó adatbázist használtam.

4.1. ábra: Legelőre alapozott állattartás a *Limousin* törzstenyészetben

A tejösszetevők becslésére egy hitelesített laboratórium bocsátott rendelkezésemre adatokat a modell feltanítására. Ez az adatbázis három farmról származott és 3 év adatát tartalmazta (2019–2021). Az adatbázis felhasználás előtt anonimizálásra került, azaz sem a tenyésztők, sem a laboratórium érzékeny adatait nem tartalmazta. Az eredeti adatbázis a Magyar Agrár és Élettudományi Egyetem Műszaki Intézetében elérhető. A tenyésztők havonta küldik a tejmintákat ellenőrzésre a laboratóriumba, így tenyésztőnként 36 havi adatot tartalmaz az adatbázis. Összesen 25 000 mérés volt és mindegyik mérés 14 különböző tejparaméter-értéket tartalmazott. A 14 paraméterből nem mindegyiket használtam fel az egyes modellekhez. Így az adatbázis lehetőséget biztosított többféle modell felállítására is attól függően, hogy mit választok kimeneti változónak. Kutatásom során a szomatikus sejtszám és a kazeintartalom becslésével foglalkoztam, mivel ezek a legfontosabb paraméterek a tenyésztők számára. A szomatikus sejtszámból a tehén egészségi állapotára lehet következtetni, a kazein pedig nagyban befolyásolja a tej minőségét. Így mindkét modell hasznos lehet a tenyésztők számára, mert áralakító tényező.

4.1. Módszerek

A modellek építéséhez, tanításhoz többféle gépi tanulási módszert használtam.
A gépi tanulásban az algoritmusok azok a matematikai vagy statisztikai modellek és módszerek, amelyek lehetővé teszik, hogy a mesterséges intelligencia tanuljon az adatokból, és ennek alapján következtetéseket vonjon le.

Az algoritmusok lényege, hogy segítsenek a minták felismerésében, a becslések készítésében és a problémák megoldásában. Az algoritmusok problémák és adatok szerint is más-más célra használhatók (optimálisak), és általában három nagy kategóriába sorolhatók.

Felügyelt tanulás (Supervised Learning)
Az ilyen algoritmusok egy tanuló adathalmazzal dolgoznak, amely tartalmazza a bemeneti adatokat és a hozzájuk tartozó címkéket vagy eredményeket. Az algoritmus célja, hogy megtanulja a bemenetek és a kimenetek közötti összefüggést, majd új adatokon pontos előrejelzéseket adjon.

A leggyakrabban használt algoritmusok:

- *Lineáris regresszió*: Folytonos kimeneti változókat (pl. ár, hőmérséklet) becsül meg.
- *Logisztikus regresszió*: Bináris osztályozási problémákra használják (pl. betegség jelenléte vagy hiánya).
- *Döntési fák* (Decision Trees): Adatok hierarchikus elágaztatása alapján hoz döntéseket.
- *Tartóvektorgépek* (Support Vector Machines, SVM): Az adatokat egy síkkal választja el, amely maximalizálja az osztályok közötti távolságot.
- *Neurális hálózatok*: Összetettebb mintázatok felismerésére és mélytanulásra használatosak.

Felügyelet nélküli tanulás (Unsupervised Learning)
Itt az adathalmaz nincs címkézve, azaz az algoritmus nem tudja, hogy az adatokhoz milyen kimenet tartozik. A cél az, hogy az algoritmus az adatokban meglévő rejtett mintákat vagy szerkezeteket találja meg.

A leggyakrabban használt algoritmusok:

- *K-közép algoritmus* (K-Means): Az adatokat csoportokra (klaszterekre) bontja.
- *Főkomponens-analízis* (Principal Component Analysis, PCA): Dimenziócsökkentésre használják. Az adatokat így kevesebb változóval (főkomponensekben) lehet jellemezni, ami gyorsabb és egyszerűbb elemzést tesz lehetővé.

- *Hierarchikus klaszterezés*: Az adatok hierarchikus struktúrában történő csoportosítása.

Megerősítéses tanulás (Reinforcement Learning)
Ez a módszer egy ágens (agent) tanulási folyamatán alapul, amely különböző környezetekben működik. Az ágens akciókat hajt végre, és visszacsatolásként jutalmat vagy büntetést kap a cselekedeteiért. A cél, hogy az ágens megtanulja, hogyan maximalizálja a jutalmat.

A leggyakrabban használt algoritmusok:

- *Q-learning*: Az ágens értékeli az egyes lépések várható értékét különböző állapotokban.
- *Deep Q-Networks* (DQN): A mély neurális hálókat használják a megerősítéses tanuláshoz.
- *Policy gradient módszerek*: Közvetlenül a cselekvések valószínűségét optimalizálják.

A megfelelő algoritmus kiválasztásánál többféle szempontot is figyelembe kell venni: struktúrált vagy strukturálatlan a rendelkezésre álló adathalmaz, illetve az adathalmaz mennyisége is szempont. A modell célja tehát az előrejelzés (regresszió), az osztályozás, a klaszterezés vagy a döntéshozatal. Különösen nagy adathalmazok esetén a pontosság és a teljesítmény között is kell döntenünk.
A kutatásom során több különböző algoritmust is használtam a modellek tanításához. Ezzel egyrészt össze tudtam hasonlítani az egyes modellek teljesítményét, hogy megtaláljam a legjobb eredményt mutató modellt; másrészt az egyes célok más jellegű algoritmus esetén hozták a legjobb eredményt.

A következőkben áttekintem a kutatásom során használt algoritmusokat.

4.1.1. Lineáris regresszió

A lineáris regresszió célja, hogy egy folytonos kimeneti változót (függő változó, y) előre jelezzen a bemeneti változók (független változók, X) lineáris kombinációja alapján. Ez az egyik legegyszerűbb és leggyakrabban használt regressziós módszer.

A többváltozós lineáris regresszió matematikai modellje az alábbi egyenlettel írható fel:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ij} + \varepsilon_i \,, \tag{4.1}$$

ahol:

- y_i a kimeneti változó i-edik megfigyeléshez tartozó értéke;
- x_{ij} az i-edik megfigyelés j-edik független változója (független változók, amelyek alapján az előrejelzést végezzük);
- β_j a j-edik független változóval társított együttható (megmutatja, hogy az egyes független változók mennyire befolyásolják a függő változót);
- β_0 konstans tag (metszéspont), az az y érték, amikor minden j-re $x_j = 0$;
- ε_i a hiba, amely azt mutatja, hogy a modell nem tökéletesen írja le az adatokat (általában normális eloszlású, $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$).

Az algoritmus célja a paraméterek (β_0, β_1, …, β_p) meghatározása úgy, hogy a predikciós hibát minimalizáljuk. Ezt általában az *összesített négyzetes hiba* (Sum of Squared Errors, SSE) minimalizálásával érjük el:

$$\mathrm{SSE} = \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right)^2 \tag{4.2}$$

A legkisebb négyzetek módszerével a paraméterek zárt alakban is meghatározhatók:

$$\boldsymbol{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \tag{4.3}$$

ahol:

- $\mathbf{X}$ az $n \times (p + 1)$-es mátrix, amely tartalmazza az x értékeket és az 1-eseket;
- y az n-es vektor, amely az y értékeket tartalmazza;
- β a becsült paraméterek vektora.

A gépi tanulási modellekben általában mátrix dekompozíciót vagy numerikus optimalizációt használunk (pl. gradient descent). A gradiens módszer egy iteratív optimalizációs algoritmus többváltozós lineáris regresszió és mélytanulási hálózatok paramétereinek frissítésére. Az algoritmus célja a veszteségfüggvény (loss function) minimalizálása, azaz olyan paramétereket találni, amelyek minimalizálják az adott modell hibáját.

A gradiens egy többváltozós függvény lejtésének irányát mutatja. A gradiens módszer ezt az irányt használja fel arra, hogy lépésről lépésre közelebb kerüljön a minimumhoz.

Ha egy $J(\beta)$ veszteségfüggvényt minimalizálni szeretnénk, akkor az optimalizáció a következő képlet szerint történik:

$$\beta^{(t+1)} = \beta^{(t)} - \alpha \nabla J(\beta^{(t)}) \,, \tag{4.4}$$

ahol:

- β a paramétervektor, amit frissítünk;
- α tanulási ráta (learning rate), amely meghatározza, hogy milyen nagy lépéseket teszünk;
- $\nabla J(\beta)$ a veszteségfüggvény gradiensvektora, amely az adott paramétereknél megmutatja, hogy merre kell csökkenteni a függvényt.

A gradiens előjele mindig azt az irányt mutatja, ahol a függvény nő, ezért, ha ebből kivonunk, akkor a csökkenés irányába haladunk.
A lineáris regresszió célfüggvénye a négyzetes hibaösszeg (Mean Squared Error, MSE):

$$J(\beta) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - \widehat{y_i})^2] \tag{4.5}$$

Az egyenletben:

$$\widehat{y_i} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \tag{4.6}$$

ahol:

- y_i a valós célérték;
- $\widehat{y_i}$ a modell által becsült érték;
- x_{ij} a független változó értéke;
- n az adatpontok száma.

A gradiens kiszámításához a veszteségfüggvény gradiensét az egyes paraméterekre nézve deriváljuk:

$$\frac{\partial J}{\partial \beta_j} = -\frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y_i}) x_{ij} \tag{4.7}$$

Ez megmutatja, hogy egy adott paraméter (β_j) miként változtatható a hibacsökkentés érdekében.

A gradiens módszer minden iterációban frissíti a paramétereket az alábbi módon:

$$\beta_j^{(t+1)} = \beta_j^{(t)} - \alpha \cdot \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)x_{ij} \tag{4.8}$$

ahol:

- α a tanulási ráta (learning rate);
- $\beta_j^{(t+1)}$ az új paraméter az iteráció után;
- $\beta_j^{(t)}$ az aktuális paraméter;
- $\frac{\partial J}{\partial \beta_j}$ a veszteségfüggvény meredeksége (gradiens).

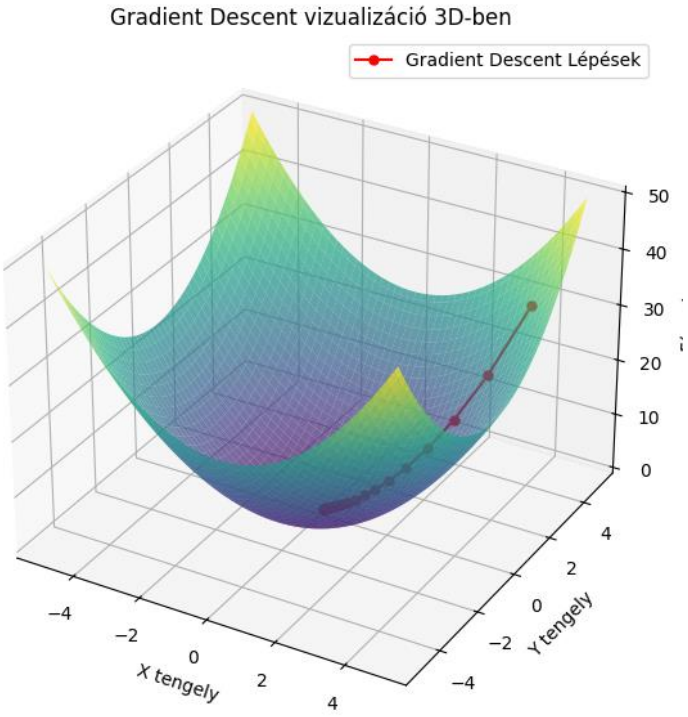
A gradiens módszer lépéseit vizuálisan a *4.2. ábrán* láthatjuk. Az ábrán egy háromdimenziós felület látható, amely egy kvadratikus függvényt reprezentál. Ez a felület a veszteségfüggvény (loss function) egy adott példája, amelyet a gradiens módszerrel optimalizáltam.

A felületet színes háló ábrázolja. Ez a kvadratikus függvény paraboloid alakú domborzatot képez. Az értékek a magasság (*z* tengely) szerint nőnek, tehát a cél az, hogy „lefelé", a minimum felé haladjunk.

A piros pontok (gradient descent lépések) láncolatot alkotnak a felületen. Ezek mutatják az iterációkat, vagyis, hogy hogyan mozog a gradient descent az induló ponttól a minimum felé. Minden egyes piros pont egy lépés a gradiens módszer során. Az *x* és *y* tengelyen a két optimalizálásra kerülő változót láthatjuk. A *z* tengely maga a veszteség.

A módszer a meredekebb részeken nagyobb lépéseket tesz, mert itt a gradiens nagyobb értékeket ad. Ahogy a pontok közelednek a minimumhoz, a lépések egyre kisebbek lesznek, mivel a gradiens lecsökken. A végén a piros pontok összegyűlnek a globális minimum körül, ahol a függvénynek legkisebb az értéke.

Ha túl nagy lenne a tanulási ráta (α), a piros pontok kaotikusan mozognának, és akár nem is érnék el a minimumot. Túl kicsi tanulási ráta esetén a lépések lassúak lennének, és az algoritmus túl sok iterációt venne igénybe.



4.2. ábra: Háromdimenziós felület a veszteségfüggvény (loss function) ábrázolására

A lineáris regresszió kiértékelésére az alábbi mutatókat használtam:

R^2 (determination coefficient – determinációs együttható)

A determinációs együttható azt méri, hogy a független változók milyen jól magyarázzák a célváltozó varianciáját. Az értéke 0 és 1 között van:

- $R^2 = 1 \rightarrow$ A modell tökéletesen magyarázza az adatokat.
- $R^2 = 0 \rightarrow$ A modell nem jobb, mint egy véletlenszerű becslés (pl. az átlag).
- $R^2 < 0 \rightarrow$ A modell rosszabb, mint egy konstans érték becslése.

Kiszámítása:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}, \tag{4.9}$$

ahol:

- y_i a valós célértékek;
- $\hat{y}$ a modell által becsült értékek;
- $\bar{y}$ a célváltozó átlaga.

MSE (Mean Squared Error – Négyzetes hiba átlaga)
Az átlagos eltérés a modell által becsült és a valós értékek között.

Egyenlete:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_1)^2] \,, \tag{4.10}$$

ahol:

- n az adatpontok száma.

Minél kisebb az MSE, annál jobb a modell illeszkedése.

Adjusted R^2 (Módosított determinációs együttható)
A módosított R^2 figyelembe veszi a változók számát. Ha több prediktort adunk a modellhez, az R^2 mindig nőhet, de a módosított R^2 kiszűri a fölösleges változókat.

Egyenlete:

$$\mathrm{R}^2_{\mathrm{adj}} = 1 - (1 - \mathrm{R}^2) \times \frac{n-1}{n-p-1}, \tag{4.11}$$

ahol:

- n az adatpontok száma;
- p a független változók száma.

4.1.2. Poisson-regresszió

A Poisson-regresszió olyan speciális regressziós modell, amelyet diszkrét, nemnegatív egész számú kimeneti változók (pl. eseményszám) modellezésére használnak. Tipikus példa a küllemi bírálati pontozás eredménye (mert néhány diszkrét értéket vehet fel).
A Poisson-regresszió matematikai modellje:

$$y_i \sim \mathrm{Poisson}(\lambda_i), \quad \lambda_i = \exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\right), \tag{4.12}$$

ahol:

- y_i Az i-edik megfigyeléshez tartozó eseményszám;
- λ_i Az események várható értéke az i-edik megfigyelésnél;
- β_j A j-edik változóval társított paraméter;
- a várható értéket (λ) a *logaritmus kapcsolat* ($\log(\lambda)$) szabályozza.

Az algoritmus célja a paraméterek becslése maximum likelihood (ML) módszerrel. A Poisson valószínűségi tömegfüggvény:

$$P(y_i|\lambda_i) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} \tag{4.13}$$

A becslés során a cél a log-likelihood (legnagyobb valószínűség) függvény maximalizálása:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i \, log(\lambda_i) - \lambda_i - log(y_i!)] \tag{4.14}$$

4.1.3. Random Forest

A *Random Forest* egy ensemble (együttes tanulási) módszer, amelyet osztályozási és regressziós problémákra is használhatunk. A döntési fákra épül, de több fát („erdőt") használ, és ezek eredményeit átlagolja (regresszió esetén) vagy szavazással egyesíti (osztályozási feladatoknál), hogy robusztusabb és pontosabb eredményt érjen el.
Az algoritmus véletlenszerűen kiválasztott adatmintákon dolgozik. Ezeket a mintákat visszatevéses mintavétellel (bootstrap sampling) állítja elő az eredeti adathalmazból.
Egy döntési fa egy ilyen mintán fog tanulni. Így minden fa más-más részhalmazt lát az adatokból. Mivel minden döntési fa különböző adathalmazból tanul, a fa építése során az algoritmus nem az összes elérhető bemeneti változót (jellemzőt) vizsgálja meg, hanem csak egy véletlenszerűen kiválasztott változóhalmazból választ (split). Ez a randomizáció segít csökkenteni a korrelációt az egyes fák között.
Regresszió esetén az egyes fák előrejelzéseit átlagoljuk. Az eredmény a végső becslés (*M* a fák száma):

$$\hat{y} = \frac{1}{M} \sum_{m=1}^{M} \widehat{y^{(m)}} \tag{4.15}$$

Osztályozás esetén minden fa „szavaz" a maga által becsült osztályra, és a legtöbb szavazatot kapott osztály lesz a végső döntés (többségi szavazás).

Az algoritmus működése:

1. *Mintavétel:*
    - Az eredeti adathalmaz $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
    - Készítsünk M darab bootstrap mintát: $D_1, D_2, \ldots, D_M$, ahol minden mintában az n-es elemszám megtartása mellett ismétlődhetnek adatok.
2. *Döntési fa építése:*
    - Egy adott fánál az osztási szabály kiválasztásakor csak egy véletlenszerűen választott m számú változóhalmazból (feature subset) választjuk ki a legjobb osztást.
    - Például, ha az adathalmaznak 100 változója van, és $m = 10$, akkor az osztási szabály kiválasztásához csak 10 változót vizsgálunk.
3. *Végső előrejelzés:*
    - Az összes döntési fa által adott eredményekből számoljuk az átlagot (regresszió) vagy a többségi szavazatot (osztályozás).

A Random Forest algoritmusnak többféle előnye is van. A randomizáció miatt az egyes fák nem azonos mintákon és változókon tanulnak, így kevésbé valószínű, hogy túltanulják az adathalmazt. Nem feltételez lineáris kapcsolatot a bemeneti változók és a célváltozó között.
A Random Forest algoritmus képes megmutatni, hogy mely bemeneti változók járultak hozzá leginkább a modell pontosságához (feature importance). Ez a tulajdonsága jól használható további modellek kereséséhez is.

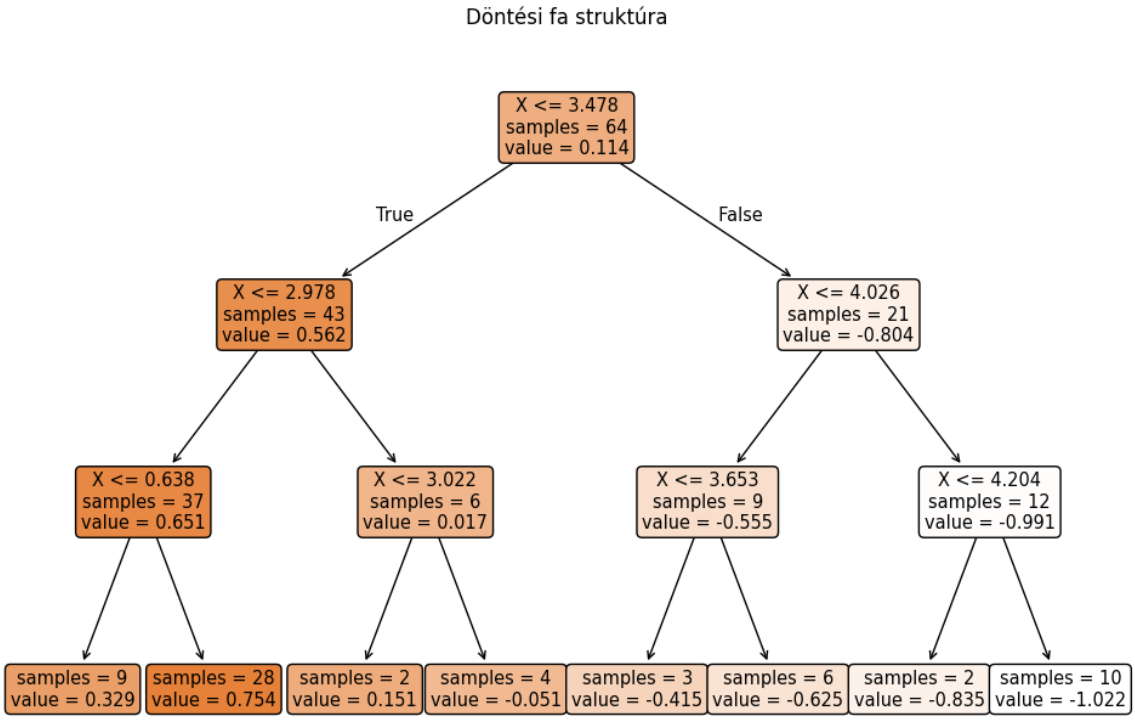
4.1.4. Decision Tree Regressor

A *Decision Tree Regressor* (döntési fa regresszió) egy fa-alapú modell, amelyet általában folytonos célváltozók előrejelzésére használnak. Ez az algoritmus a döntési fa (decision tree) működését követi, de a kimeneti változó egy numerikus érték, nem pedig egy kategória. A döntési fa regresszió az adathalmazt rekurzív módon felosztja kisebb szegmensekre úgy, hogy minden egyes osztásnál minimalizálja az adott csomóponton belüli varianciát vagy egyéb hibaértéket. Az így létrejövő fa leveleiben a célváltozó átlagértékét tároljuk.
Az algoritmus egy olyan változót és küszöbértéket keres, amely a legjobb osztást eredményezi.

A döntési fák osztási mérőszámai között a négyzetes hiba kiszámításával tesz különbséget.

Az algoritmus minden egyes osztás után az újonnan létrejött két részhalmazra ugyanezt az eljárást alkalmazza (rekurzív osztás). A fa mélysége nő, ahogy az adatok egyre kisebb csoportokra oszlanak. Az osztások addig folytatódnak, amíg a levélcsomópontban lévő minták száma egy adott küszöbérték alá nem csökken vagy egy (előre meghatározott) maximális famélységet el nem érünk.

Egy új adatpont becslésekor az algoritmus végighalad a fán a bemeneti változók értékei alapján, míg egy levélcsomóponthoz nem ér. Az előre jelzett érték a levélben található adatok célváltozójának átlaga. Az algoritmus előnye, hogy könnyen értelmezhető és vizualizálható, nem igényel normalizált adatokat, kezelni tudja a nemlineáris kapcsolatokat. Ez a módszer képes automatikusan kiválasztani a legfontosabb bemeneti változókat, ami nagy segítség a több bemeneti változót tartalmazó adathalmazok esetében. Vigyázni kell azonban, mert hajlamos túlilleszkedésre (overfitting), ha nem szabályozzuk megfelelően a fa mélységét. A döntési fa egy lehetséges felépítését a *4.3. ábrán* láthatjuk.



4.3. ábra: A Decision Tree Regressor algoritmus működése

4.1.5. Extra Trees Classifier

Az *Extra Trees (Extremely Randomized Trees) Classifier* egy döntési fa alapú algoritmus, amely több fát tanít egy adathalmazon, majd az egyes fák előrejelzéseit egyesíti a végső döntés meghozatalához. Használatának fő célja, hogy csökkentse a varianciát (overfitting elkerülése, ebben hasonlít a Random Forest algoritmushoz) és hogy nagy adathalmazokon gyorsabb legyen, mint a Random Forest. A Random Forest algoritmus minden fát egy újramintavételezett mintán tanít, ezzel szemben az Extra Trees a teljes adathalmazt használja minden fa tanításához.

A Random Forest algoritmus minden egyes döntési fa csomópontjában véletlenszerűen kiválaszt egy jellemzőhalmazt, majd ezek közül a legjobb elválasztó pontot választja ki a csomópont tisztaságának maximalizálása érdekében. A tisztaság itt azt jelenti, hogy a csomópontban lévő adatok mennyire homogének, vagyis mennyire hasonlóak egymáshoz. Ez a módszer segít abban, hogy a különböző fák eltérő felosztásokat hozzanak létre, és így az együttes modell egy robusztusabb és általánosabb előrejelzést adjon. Az Extra Trees ezzel szemben véletlenszerű osztási pontokat választ a lehetséges értékek közül, és az így kapott osztások közül választja a legjobbat. Az algoritmus több független fát épít (például 100-at), és az osztályozásnál a többségi szavazás dönt.

4.1.6. Support Vector Machine

A *Support Vector Machine* (SVM**)** egy felügyelt tanulási algoritmus, amelyet osztályozási és regressziós feladatokra egyaránt használhatunk. Osztályozási feladatnál a tanítópontokat kell két (vagy több) osztályra osztanunk. Optimális lineáris szeparálás esetében az elválasztó egyenes (sík, hipersík) a két osztályba tartozó tanítópontok között a pontoktól a lehető legnagyobb távolságra helyezkedik el. Az elválasztó felületet a tanítópontoktól egy biztonsági sáv választja el (margó vagy margin). A gyakorlatban legtöbbször az egyes osztályok nem szeparálhatók lineárisan úgy, hogy még osztályozási tartalék is biztosítható legyen. Általában a két osztály között csak nagyon kis biztonsági sáv alakítható ki, vagy a lineáris szeparálás hibamentesen nem is lehetséges. Ha nem ragaszkodnánk ahhoz, hogy minden tanítópont a sávon kívül helyezkedjen el, akkor a biztonsági sáv növelhető lenne. A sáv mérete jelentősen növelhető úgy, hogy néhány mintapontnál megengedjük, hogy a sávon belülre kerüljenek.

Az SVM célja éppen ez, hogy az egyes osztályokat ne egy diszkrét érték alapján válassza el egymástól, hanem egy olyan hipersík megtalálása, amely jobb eredményt ad, mint a diszkrét választóvonal.
A hipersík egy N-dimenziós térben egy (N−1)-dimenziós sík, amely elválasztja az osztályokat.
Az algoritmus névadói azok a tartóvektorok (Support Vectors), amelyek a legközelebb vannak a döntési határhoz, és így meghatározzák annak helyzetét. A margó a két osztály legközelebbi pontjai közötti távolság – az SVM célja ennek maximalizálása.

Az SVM működése lineáris esetben:
Adott egy (x_i, y_i) tanítóhalmaz, ahol:

- x_i az i-edik minta;
- $y_i \in \{-1, 1\}$ a hozzárendelt címke.

A cél az, hogy egy olyan hipersíkot találjunk, amely az egyes osztályokat maximális margó mellett választja el:

$$w^T x + b = 0\,, \tag{4.16}$$

ahol:

- w a normálvektor (irányvektor);
- b az eltolás (bias).

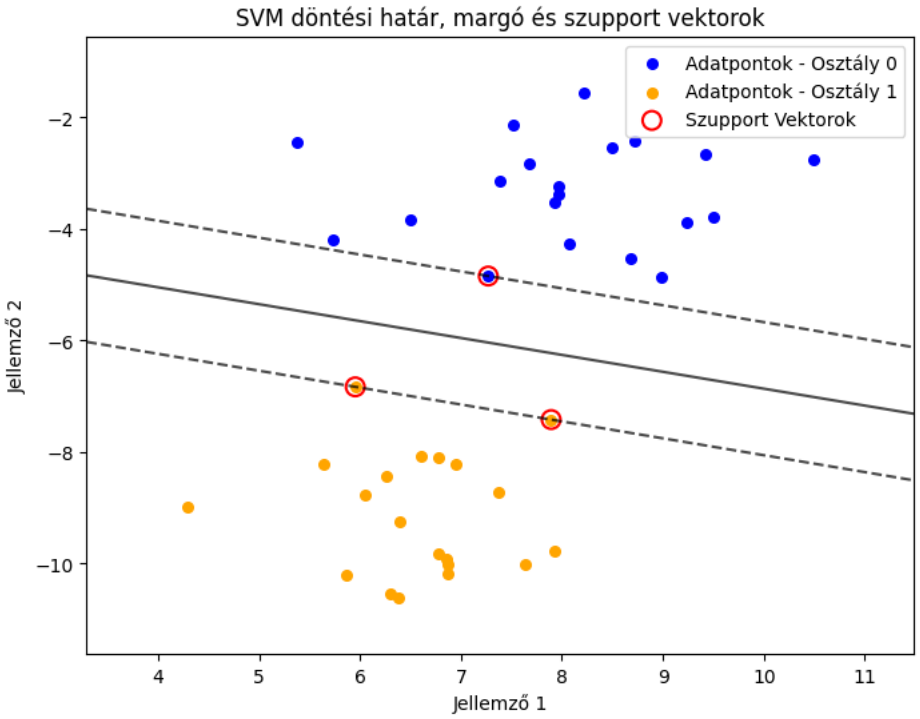
Az optimális margó maximalizálása az alábbi feltételekkel történik:

$$y_i(w^T x_i + b) \geq 1, \quad \forall i \tag{4.17}$$

A cél a következő egyenlet optimalizálása:

$$min\frac{1}{2}\left|\left|w\right|\right|^2 \tag{4.18}$$

Az SVM működését a *4.4. ábra* mutatja be.



4.4. ábra: Az SVM algoritmus működése

4.1.7. Döntési fák kiértékelési módszerei

Pontosság (Accuracy)

Azt méri, hogy a modell hány esetben adott helyes predikciót az összes minta közül.

$$\text{Pontosság} = \frac{\text{Helyes osztályozások száma}}{\text{Összes minta száma}}$$

Pozitív pontosság (Precision)

Az összes pozitívnak jelölt minta közül hány volt valóban pozitív.

Pozitív emlékezet (Recall)

Az összes valódi pozitív minta közül hányat sikerült megtalálni.

4.1.8. Hagyományos matematikai módszerek

A gépi tanulásos regressziós modellel végzett feladatokat a hagyományos matematikai módszerekkel is elvégezhetjük. Kutatásaim során a modelljeim teljesítményértékelése miatt hagyományos matematikai módszerekkel is ellenőriztem az eredményt.

Ehhez a Python *statsmodels* nyílt forráskódú modulját használtam. Ezt a modult többszörösen validált matematikai statisztikai módszerek számítógépes szimulálására

fejlesztették ki (*Seavold et al.,* 2010). A *statsmodels* könyvtár OLS (Ordinary Least Squares) függvényét használtam a kutatásaim eredményének matematikai modellekkel való összehasonlítására.

A matematikai módszerek hátránya elsősorban nagyméretű adatbázisok esetében jelenik meg, ezért van előnye, ha gépi tanulásos módszerrel is tudunk hasonló vagy jobb modellt építeni.

Az OLS mátrixinverziót használ a megoldás során, amely $O(n^3)$ időkomplexitású, így nagy adathalmazoknál nagyon lassú.

$$\boldsymbol{\beta} = (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{y} \tag{4.19}$$

A mátrixinverzió számítása gyakorlatilag lehetetlen a bemeneti mátrix bizonyos mérete fölött.

Bár a mátrixinverzió elméletileg bármilyen méret esetén elvégezhető, 500×500-as vagy annál nagyobb mátrixoknál a művelet túl lassú, vagy numerikusan instabil (kerekítések miatt hiba keletkezik), ezért a gyakorlatban ritkán használják. Ezzel szemben a gépi tanulás során a gradiens ereszkedés iteratív módszerként működik, $O(n)$ komplexitású lépésekkel, ezért sokkal gyorsabb nagy adatállományoknál. Ezenkívül az OLS egy statikus modell, azaz egyszer lefuttatjuk, és nem tanul új adatokból. Ezzel szemben a gépi tanulás lehetővé teszi az online tanulást, tehát új adatok érkezésekor a modell folyamatosan, automatikusan frissíthető. Ha az adatok túl sok változót tartalmaznak, és multikollinearitás lép fel, az OLS rossz becsléseket adhat, ilyenkor jellemzően túlilleszkedés lép fel. A gépi tanulás során regularizációval csökkenthetjük a túlilleszkedést.

4.1.9. Kiegyensúlyozatlan adathalmazok

A valós környezetben a legrosszabb kimenetű adatok száma általában kisebb – máskülönben a meglévő tenyésztési (vagy gyártási) módszerek teljesen életszerűtlenek lennének –, így valóságban gyakran *kiegyensúlyozatlan adathalmazzal* találkozunk a historikus adatok feldolgozásánál. Mivel kutatásom során a tejmintákban a beteg állatok mintáit fogom keresni és a betegséget tervezem előre jelezni, várható, hogy erősen kiegyensúlyozatlan lesz az adatbázis. Ahhoz, hogy megfelelő modellt készíthessek, mindenképpen ki kell egyensúlyoznom az adathalmazt, máskülönben nagyon elfogult, pontatlan modellt kapnék.

A kiegyensúlyozatlan adathalmazok kiegyensúlyozására (adatszintű megközelítésben) alapvetően kétféle módszert alkalmazhatunk. Az egyik módszer a túlreprezentált adatok elhagyása és további alulreprezentált adatok beszerzése. A másik módszer az alulreprezentált adatok többszörözése (SMOTE – Synthetic Minority Over-sampling Technique). Ezekkel a módszerekkel a modell pontossága sérülhet, de az érzékenysége általában javul.

A módszerek hatékonyságát és esetleges további javítását szintetikus adatbázisok létrehozásával és ezek elemzésével ellenőriztem.

Első lépésben *Python* program segítségével létrehoztam egy szintetikus adatbázist. Ehhez a Python *Pandas* és *Numpy* moduljait használtam. Az adatbázist úgy készítettem el, hogy négy bemeneti változója (folytonos változók) és egy kimeneti változója legyen. A kimeneti változót úgy állítottam elő, hogy kategorikus változó legyen háromféle kimeneti értékkel, osztállyal (0, 1 és 2). Az adatbázis 2500 mintát tartalmazott.

A kiinduló adatbázisban a kimeneti változók eloszlása a *4.1. táblázatban* látható.

4.1. táblázat: A szintetikus adatbázis kezdeti eloszlása

Osztály	**Arány**
0	45,8%
1	44,6%
2	10%

Mindkét kiegyenlítési módszerrel kiegyenlítettem az adatbázist, majd Random Forest predikciós algoritmussal megvizsgáltam az osztályozás precizitását (precision), és a fedést (recall) is. Ezeket az értékeket minden kimeneti osztályra megvizsgáltam, majd vettem az átlagukat, és ezt tekintettem a módszer eredményességének. Az eredmények a *4.2. táblázatban* láthatók.

4.2. táblázat: A kétféle kiegyenlítési módszer összehasonlítása

Módszer	**Precizitás**	**Fedés**
Véletlen túlmintavételezés	0,36	0,35
SMOTE	0,34	0,34

Látható, hogy mindkét módszer hasonló eredményt hozott. Egy valós adatbázisnál elképzelhető, hogy tudunk újabb adatokat bevonni a tanító adatbázisba (de nem eleget), ezért megnéztem, hogy a két módszer vegyes alkalmazásával milyen eredményt lehetne elérni *(4.3. táblázat)*.

4.3. táblázat: A hibrid kiegyenlítési módszer predikciós eredményei

Módszer	**Precizitás**	**Fedés**
Hibrid	0,36	0,36

Látható, hogy az eredmény kismértékben javult, hiszen mind a fedés, mind a precizitás értéke nőtt.

Megvizsgálva a kutatásom során használt adatbázisokat, azt tapasztaltam, hogy azonos kimeneti osztály esetén is a bemeneti változók értékei kismértékben eltérhetnek egymástól. Ezért a modell érzékenységét tovább növelheti, ha a fenti módszerekkel beszúrt új sorok esetében a bemeneti változók értékét egy kicsit módosítom. Ezért a bemeneti változókat 0,98–1,02 közötti véletlenszerű értékkel megszoroztam, és így végeztem el ismét az eredményességi vizsgálatot (adataugmentáció).

5. EREDMÉNYEK

A kutatásom során olyan feladatokra kerestem megoldás, amelyekre korábban nem alkalmazták a gépi tanulással feltanított modelleket. A szarvasmarha küllemi bírálatokhoz és a tejtermeléshez kapcsolódó adatbázisok felkutatásával értékes adatbázisok álltak rendelkezésemre.

Három esetben bizonyítottam be, hogy a gépi tanulás segítségével készíthető olyan rendszer, amely akár a mindennapi életben is jól használható. Ezek alapján megvizsgáltam a szarvasmarha küllemi bírálati pontszámok becslésének lehetőségét. Ezenkívül laboratóriumi tejmintákban a szomatikus sejtszám, valamint a kazein mennyiségének a becslését végeztem el.

Kutatásom során bebizonyítottam, hogy készíthető megfelelő, gépi tanuláson alapuló modell ezekre a felhasználási esetekre, és ezek a modellek jobban teljesítenek a hagyományos matematikai modelleknél.

5.1. Küllemi bírálati pontok becslése gépi tanulás segítségével

A küllemi bírálati pontozás fontos eszköz a szarvasmarha-tenyésztésben, amely segíti a tenyésztőket a legjobb genetikai adottságokkal rendelkező egyedek kiválasztásában. A megfelelő testfelépítés és tőgyszerkezet nemcsak a termelésben, hanem az állat egészségének megőrzésében és élettartamának növelésében is kulcsszerepet játszik. Ezáltal a küllemi bírálat hozzájárul egy gazdaságilag hatékonyabb és fenntarthatóbb szarvasmarha-tenyésztési rendszer kialakításához.

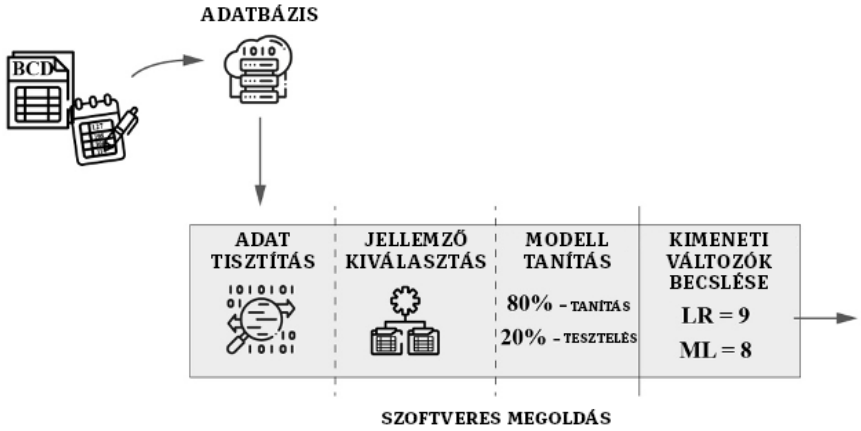
5.1. ábra: Szarvasmarha küllemi bírálat

5.1.1. Küllemi bírálati pontszámok becslése

Kutatásom során kétféle gépi tanulási módszer eredményességét hasonlítottam össze a szarvasmarhák küllemi bírálati pontjainak becsléséhez: a *lineáris regresszió* és a *Poisson-regresszió* alapú modell teljesítményét vetettem össze a szarvasmarhák különböző fenotípusos paramétereinek előrejelzésére. Azt szerettem volna megtudni, melyik modell ad jobb eredményt, és hogyan viszonyulnak ezek az eredmények a hagyományos matematikai modellekhez. Az egyes modellek teljesítményének összehasonlítására a négyzetes hibát és a determinációs együtthatót (R^2) használtam. Az eredmények ismeretében kijelenthetjük, hogy készíthető-e MI-alapú modell a szakértői becslések alapján.

A modell tanításához használt adatbázis 325 állat küllemi bírálati pontszámait tartalmazta egy Limousin törzstenyészetből, amelyet a *Limousin és Blonde d'Aquitaine Tenyésztők Egyesülete* (Budapest) állított össze. Az állatok 18 bika utódai voltak, amelyek 1990 és 1996 között születtek. A 12 hónapos állatokat hivatalosan is minősítették a teljesítményvizsga végén.

A kutatás során egy olyan rendszer kialakítása volt a célom, amely később könnyen újra használható más tenyésztők adatainak a felhasználásával. A predikciós rendszer felépítése az *5.2. ábrán* látható.



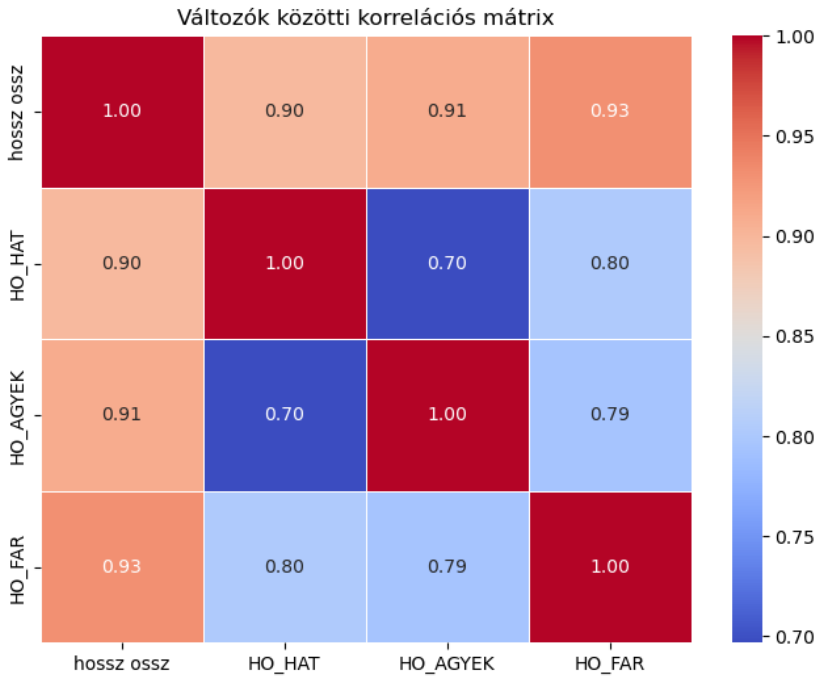
5.2. ábra: A becslő szoftver működésének elvi vázlata

Az adatbázisban tárolt korábbi értékelési adatokat *(5.1. táblázat)* használtam különböző regresszióalapú mesterséges intelligencia algoritmusok betanítására. Az egész rendszert Pythonban készítettem el. Az adattisztítást és -átalakítást szintén Pythonban végeztem. A modell tanításához a *Scikit-learn* Python-modul segítségével készült az egyedi kód. A modell betanításához az adatbázist két részre osztottam. Az adatok 90%-át használtam az algoritmus betanítására, a fennmaradó 10%-ot pedig a tanulás utáni tesztelésre és az algoritmus validálásra.

5.1. táblázat: A kimeneti változók ismertetése

Függő változók	**Független változók**
Izmoltság (0–60 pont)	Használati pontszám, hossz, szélesség (0–40 vagy 0–60 pont)
Far hossza (1–9 pont)	Testhossz, háthossz (1–9 pont)
Mellkas izmoltsága (1–9 pont)	Egyes részek izmoltsága (1–9 pont)
Far izmoltsága (1–9 pont)	Egyes részek izmoltsága (1–9 pont)

Minden változó folytonos volt. Megvizsgáltam a közöttük lévő korrelációt, amelynek eredményét az alábbi hőtérképén láthatjuk *(5.3. ábra)*.



5.3. ábra: A kiválasztott változók hőtérképe (korrelációs mátrixa)

A korrelációk 1-hez közeli értéket mutattak, ezért jó feltételezés, hogy a regressziós modellek jó eredményt fognak adni.

Minden modell tanítását 10-szer futtattam le úgy, hogy mindegyik esetben az adatbázist véletlenszerűen osztottam két részre. Ezeket az eredményeket összehasonlítottam, és végül mindegyik modell esetében a legjobb eredményt szerepeltettem a táblázatban.

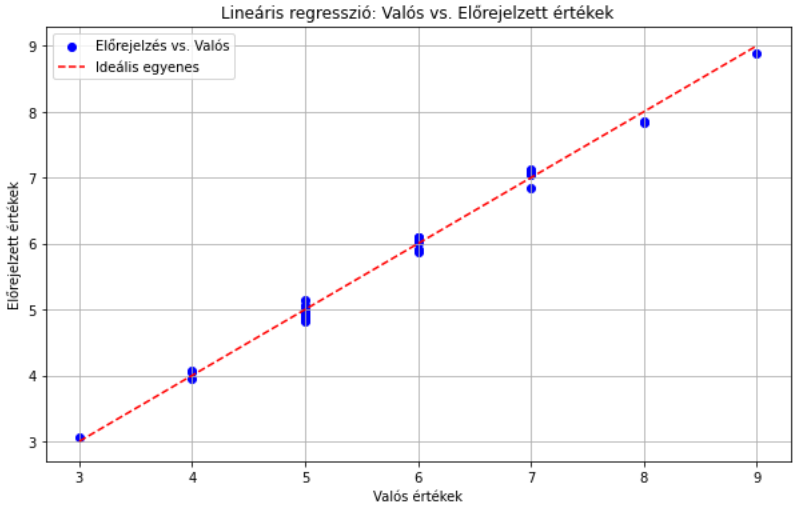
A feltanított modelleket a tesztadatokkal ellenőriztem. Az ellenőrzés eredménye a *5.2. táblázatban* látható.

5.2. táblázat: A kétféle modell eredményei (*n* = 325)

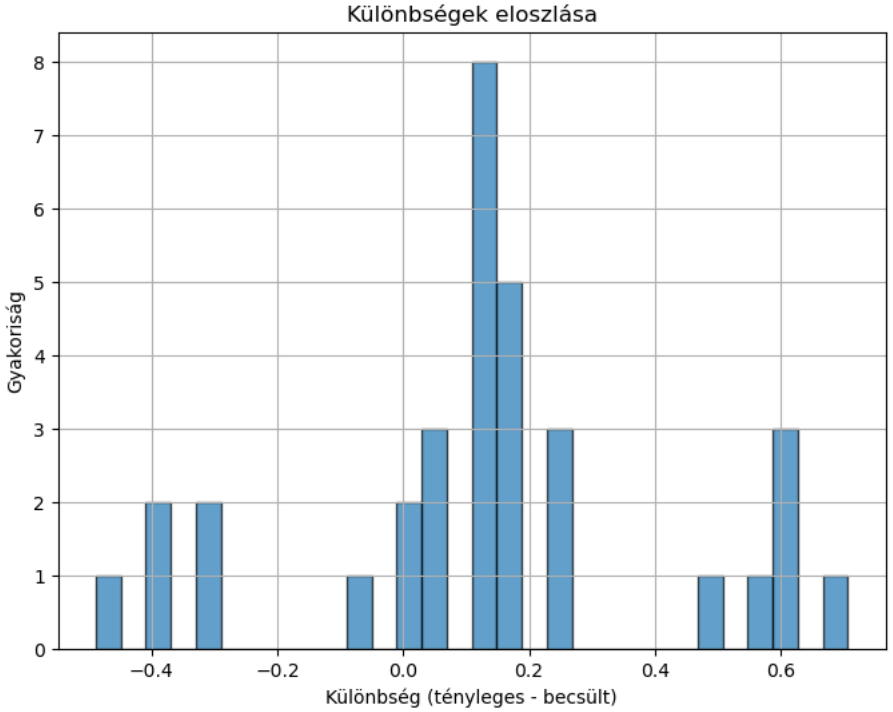
Modell típusa	**Függő változók**	**MSE**	**R^2**	**A-R^2**
Lineáris regresszió	Izmoltság (0–60)	3.38	0.86	0.83
	Far hossza (1–9)	0.21	0.93	0.91
	Mellkas izmoltsága (1–9)	0.35	0.86	0.79
	Far izmoltsága (1–9)	0.41	0.77	0.76
Poisson-regresszió	Izmoltság (0–60)	4.00	0.81	0.77
	Far hossza (1–9)	0.23	0.92	0.85
	Mellkas izmoltsága (1–9)	0.39	0.82	0.78
	Far izmoltsága (1–9)	0.49	0.73	0.7

Az eredményekből jól látható, hogy az R^2 és az A-R^2 értékek nagyon hasonlóak egymáshoz, különösen az izmoltság (Score for Muscularity) esetében. Ez azt jelenti, hogy a modellválasztás megfelelő volt, és mindkét modell jól használható a pontértékek becslésére. Azonban a lineáris regressziós modell R^2 értéke kissé jobb volt, mint a másik módszer eredménye a far hosszúságának előrejelzésénél. A mellkas izomzatának becslésénél (Muscularity of Breast) elért eredmények hasonlóak voltak mindkét modell esetében. A Poisson-regresszió alapú modell a far szélességét is rosszabb eredménnyel becsülte. A lineáris regressziós modell alkalmazása során a négyzetes hibák értékei is kicsik voltak, ezért – az eredmények alapján – ezt az algoritmust érdemes használni a jövőbeli gyakorlatban. A Poisson-regressziós módszer alkalmazása során mind a négy paraméter esetében nagyobb hibát kaptunk.

Az eredményeim kis variabilitást mutatnak, de jól tükrözik az anatómiai összefüggéseket. Nem meglepő, hogy az izomzat értekéből (Score for Muscularity) jól előre jelezhetők a többi három jellemző eredményei ($R^2 = 0,86$).
A tárolt és a becsült eredmények vizualizálásához egy egyenes vonalat húztam a mért értékekből, és ugyanabban a grafikonban kék pontokkal ábrázoltam a becsült értékeket *(5.4. ábra)*.



5.4. ábra: Eredeti tárolt értékek és a becsült értékek összehasonlítása a kimeneti változóra



5.5. ábra: A különbségek (valós érték – előre jelzett érték) eloszlásának ábrázolása, amely segít felmérni a modell hibáit

Az egyik kimeneti változó (far hosszának becslése) esetében hagyományos matematikai módszerrel (legkisebb négyzetek módszere) is készítettem egy modellt. Így összehasonlítható a gépi tanulás eredménye a hagyományos matematikai módszerekkel. A matematikai modell eredményeit a *5.3. táblázat* tartalmazza.

5.3. táblázat: OLS-eredmények

Függő változó	Far hossza	**R^2**	0.91
Model	OLS Adj.	**Megfigyelések száma**	325
Módszer	Legkisebb négyzetek		

Az eredmények jól mutatják, hogy a gépi tanuláson alapuló modellek segítségével előre jelezhetők a szarvasmarhák küllemi tulajdonságai. A gépi tanulással létrehozott modell gyakorlati alkalmazhatóságát tekintve a 86%-os érték nagyon jónak mondható. A mérési eredmények elemzése alapján elmondható, hogy ha több historikus adatot használhatunk a modell tanításához, akkor még pontosabb és általánosabb modellt alkothatunk, amelyek mindennapi segítséget nyújthatnak a gazdák számára a korai szelekcióban. A viszonylag kis elemszám (n = 325) ellenére a modell túlilleszkedésének kockázata alacsony, mivel csak 3 bemeneti paramétert használtam, és a lineáris regresszió nem hajlamos túlilleszkedésre ilyen arány mellett.

A gépi tanulás segítségével készített modell jobb eredményt hozott, mint a hagyományos matematikai modell, valamint a tanítás a meglévő szoftverkörnyezet felhasználásával automatikusan elvégezhető. További fejlesztői vagy szakértői beavatkozás nem szükséges a rendszer adaptálásához, ami szintén segíti az esetleges későbbi alkalmazást.

5.1.2. Regressziós modellek korábbi eredményei

A küllem becslésre regressziós modell segítségével korábban is történt kísérlet, ezek azonban nem gépi tanulási modellt használtak. Ezeket az eredményeket is átnéztem, hogy a gépi tanuláson alapuló modellem eredményét ezekhez az eredményekhez is tudjam viszonyítani.

Korábbi kutatásokban klasszikus matematikai modellt használtak a húshasznú marhák bizonyos jellemzőinek becslésére. Ezek az eredmények jó alapot nyújtanak arra, hogy összehasonlítsam az MI-modellemet a hagyományos matematikai modellekkel. A hagyományos statisztikai regressziók eredményei, amelyek az élősúlyra és a herezacskó kerületére vonatkoztak a szarvasmarha-tenyésztésben, az *5.4. táblázatban* láthatók, referenciaként az MI-algoritmus számára.

5.4. táblázat: Élősúly és here kerületének becslése matematikai regressziós modellekkel

Célérték	Faj	Eredmény	Forrás
Herekörméret (SC)	Különböző fajták	Az élősúly erős vagy közepes korrelációt mutatott a herezacskó kerületével.	Boudron et al. (1986), Knights et al. (1984)
Élősúly becslése	Nőstény tejelő szarvasmarha, amely főként az őshonos zebuból és Guzerattal vagy Bos Taurusszal való keresztezéséből áll.	A legjobb modell, amely megjósolta az élősúlyt a szívkerület alapján készült 0,85-ös korrigált R^2 értékkel.	Tebug et al. (2018)
	Holstein, Brown Swiss	A mellkasméret volt a legjobb paraméter a testtömeg előrejelzésére a Brown Swiss (R^2 = 91,1%) és a keresztezett szarvasmarhák (R^2 = 88,8%) esetében, összehasonlítva Holsteinnel (R^2 = 60,7%).	Ozkaya et al. (2009)
	Indigenous, Friesian, Brahman, Red Dane	Az élősúly erősen korrelált (r = 0,90) a test hosszával, a szív kerületével és a marmagassággal. A szív kerületével való korreláció nagyon erős volt (r = 0,96).	Francis et al. (2002)
	Holstein	Elegendő a szívkerület mérése (R^2 = 0,95) az élősúly megbízható előrejelzéséhez hat hónapos korig nőstény borjakban.	Ashwini et al. (2019)
	Limousin marhák	Az élősúlyt elsősorban a vállak szélessége és a marmagasság határozta meg (R = 0,74).	Tőzsér et al. (2020)

A legjobb eredményeket az élősúly előrejelzésére főként a szív- és a mellkaskörfogat határozta meg a tejhasznú szarvasmarhák esetében. A Limousin fajtában az élősúlyt két paraméter befolyásolta (a vállszélesség és a marmagasság).
Ezek az előrejelzések segítik a tenyésztőket a pontosabb szelekcióban.

5.2. Szomatikus sejtszám becslése

A tej minőségének rendszeres ellenőrzése számos okból fontos: az állatok általános egészségügyi állapotának felmérésére, a tejhozam növelésére és a tejtermékek előállítására is hatással van.
Léteznek módszerek a tej minőségének helyszíni monitorozására, azonban a legpontosabb megoldást a laboratóriumi vizsgálatok jelentik, amelyeket általában havonta végeznek. Ez a módszer azonban nem alkalmazható közvetlenül a gazdaságokban. A laboratóriumi vizsgálatok évtizedek óta szabványos monitorozási módszernek számítanak, ezért nagy mennyiségű korábbi adat áll rendelkezésre a tejminőség ellenőrzéséhez. A tej számos, az ember számára fontos tápanyagot tartalmaz, például zsírokat, fehérjéket, szénhidrátokat, különféle ásványi anyagokat és vitaminokat is (*Durgesh et al.,* 2022). Ezeknek az összetevőknek a mennyisége meghatározza a tej minőségét, valamint az állatok egészségi állapotáról is fontos információkat nyújt.
A laboratóriumok is végeznek alapvetően matematikai, statisztikai módszerekkel különböző elemzéseket, előrejelzéseket. Ezen elemzések célja az olyan minták azonosítása, amelyek esetleges egészségügyi problémát jelezhetnek az állatnál. Ezeknek az elemzéseknek a pontosítása, automatizálása rendkívül jó lehetőség a gépi tanulási algoritmusok számára, mivel feltanításukhoz elegendően nagy mennyiségű adat áll rendelkezése.
A kiinduló adatbázis három farmról származik és három év adatát tartalmazta (2019–2021). A tenyésztők általában havonta küldik a tejmintákat a laboratóriumnak ellenőrzésre, így tenyésztőnként 36 havi adatot tartalmaz az adatbázis. Összesen 25 000 mérés történt, és mindegyik mérés 14 különböző tejparaméter-értéket tartalmazott.
A kutatás ezen részének az volt a célja, hogy egy olyan, gépi tanuláson alapuló modellt készítsek, amely a többi paraméter alapján megfelelően előre jelzi az SCC értékét. Ehhez először a szomatikus sejtszám értékeinek statisztikai elemzésével kellett foglalkoznom.

A szomatikus sejtszám (Somatic Cell Count, SCC) a leggyakrabban használt mutató a tőgygyulladás (masztitisz) kimutatására a tejelő szarvasmarháknál. A szomatikus sejtszám értékének emelkedése tőgygyulladás jele lehet. Az SCC értékét más tényezők is befolyásolják, mint például az életkor, a laktációs időszak, az évszak, a stressz vagy a takarmány minősége, azonban ezek kevésbé jelentősek, mint a baktériumok által okozott tényezők. Általában fertőzöttnek tekintjük a tejet, ha az SCC értéke 250 000–300 000 sejt/ml felett van.

A pontos előrejelzés készítéséhez a szarvasmarha egyéb, fent említett paramétereit is figyelembe kell venni. Ennek a kutatásnak az volt a célja, hogy bebizonyítsam: készíthető olyan, gépi tanuláson alapuló modell, amely segítségével a tejminták SCC-tartalma jól becsülhető egyéb, tejben lévő összetevők értékével.

A modell építéséhez először a bemeneti változókat vizsgáltam meg. A gépi tanuláson alapuló modellek jobban működnek a kategorikus változókkal. Ezért ahol biológiailag indokolt volt, ott a változót kategorikus változóvá alakítottam. A laktációs napok számát három csoportba osztottam (0: $n < 100$, 1: $100 < n < 200$, 2: $n > 200$). Az ellések számából is három csoportot képeztem: 1, 2 és 3+.

A kutatáshoz egy olyan, saját fejlesztésű szoftverkörnyezetet készítettem, ahol a teljes tanítási folyamat egy csomagban elvégezhető, vagyis a bemenő adatok tisztítása, átalakítása és a teszt, valamint egy tanító adatbázis szétválasztása is része a folyamatnak. Ezzel a szoftver többször is felhasználható, további adatokkal könnyedén újratanítható, finomhangolható.

A szoftvert *Python* környezetben készítettem, alapvetően a *Pandas* és a *Scikit-learn* könyvtárak használatával.

5.2.1. Szomatikus sejtszám

A tehéntejben található szomatikus sejtek a tejtermelő sejtek és immunsejtek keverékei. Ezek a sejtek bekerülnek a lefejt tejbe is, ezért alkalmas, használható az állat egészségi állapotának monitorozására és a tejminőség ellenőrzésére is. A szomatikus sejtek jelenléte szoros kapcsolatot mutat az egyik leggyakoribb betegséggel, a tőgygyulladással (*Dohoo et al.*, 1991), mivel a szerepük a fertőzés elleni küzdelem és a megrongálódott szövetek megjavítása. A szomatikus sejtszámot szinte minden országban használják a tejelő tehenek egészségének ellenőrzésére. A szomatikus sejtszámot az egy ml tejben található sejtszámmal jellemzik. Az

adatbázisban tárolt adatokat megvizsgálva a szomatikus sejtszám változó statisztikai elemzése az *5.5. táblázatban* látható.

5.5. táblázat: Az adatbázisban található szomatikus sejtszám statisztikai jellemzői

Minták száma	26 6686
Minimum	2000
25%	50 000
50%	150 000
75%	401 000
Maximum	900 000

- Ha SCC < 100 000, akkor a tehén valószínűleg egészséges.
- Ha SCC > 100 000 de < 300 000, akkor a tehén különleges figyelmet igényel.
- Ha SCC > 300 000, akkor a tehén valószínűleg fertőzött.

Mivel az SCC értéke nagyon nagy varianciát mutat, jobb, pontosabb modellt kapunk, ha átalakítjuk ezen értékeket egy másik skálára. Erre a célra használják a linearizált szomatikus sejtszámot (*Dégen et al.,* 2018). Az SCC értékek átalakításához a következő formulát használtam (*Dabdoub et al.,* 1981):

$$SCCt = [\log_2(SCC/100000)] + 3 \qquad (5.1)$$

Ezzel a logaritmizálással 9 értéket lehet rendelni az SCC értékeihez, amely a *5.6. táblázatban* látható.

5.6. táblázat: Az SCC lineáris pontszámai

Kategória száma	**SCC**	**Kategória száma**	**SCC**
1	25 000	6	800 000
2	50 000	7	1 600 000
3	100 000	8	3 200 000
4	200 000	9	6 400 000
5	400 000	–	–

Ezeket a lineáris pontszámokat ezután három csoportba osztottam (nem fertőzött = 0, lehetséges fertőzés = 1, fertőzött = 2), és ezt a három értéket fogom használni a minták osztályozásához. Ezek a csoportok lesznek a modell kimeneti értékei, a függő változó.

5.2.2. Laktoferrin

A bemeneti változók közül nem mindegyik változóra van szükség a legjobb modell tanításához. A feleslegesen használt változók bonyolítják a modellt, növelik a pontatlanságot és a tanításhoz szükséges időt is. A megfelelő paraméterek kiválasztásához használhatunk matematikai elemző módszereket, de érdemes néha az egyéb tudományterületek tapasztalatira is hagyatkozni. A laktoferrin kapcsolatát a szomatikus sejtszámmal korábban más kísérletekben is kimutatták (*Cheng et al.,* 2007).

A laktoferrin egy vasmegkötő glikoprotein (*Cheng et al.,* 2007). Az immunrendszer védekező mechanizmusának működésében kritikus szerepet játszik, részt vesz a fertőző betegségek elleni védekezésben.

Bár a laktoferrin kapcsolata a szomatikus sejtszámmal laboratóriumi kísérletekben bebizonyosodott, a való életben ez a kapcsolat nem volt jelentős, önmagában a laktoferrin nem használható az SCC-szint emelkedésének előrejelzésére. A magas laktoferrin-szintű állatoknak csak ritkán volt masztitisze. Keresnünk kell tehát további paramétereket, amelyek a laktoferrinnel együtt jól jelzik az SCC változását.

5.2.3. A kimeneti változó elemzése

Mivel a szomatikus sejtszám becslése a célom, a linearizálás utáni eredményeket tovább vizsgáltam. A kimeneti változó egyes kategóriának az előfordulását az *5.7. táblázat* tartalmazza.

5.7. táblázat: A kimeneti változó egyes értékeinek előfordulása

Kategória	Előfordulás (db)
0	17 500
1	5 200
2	2 600

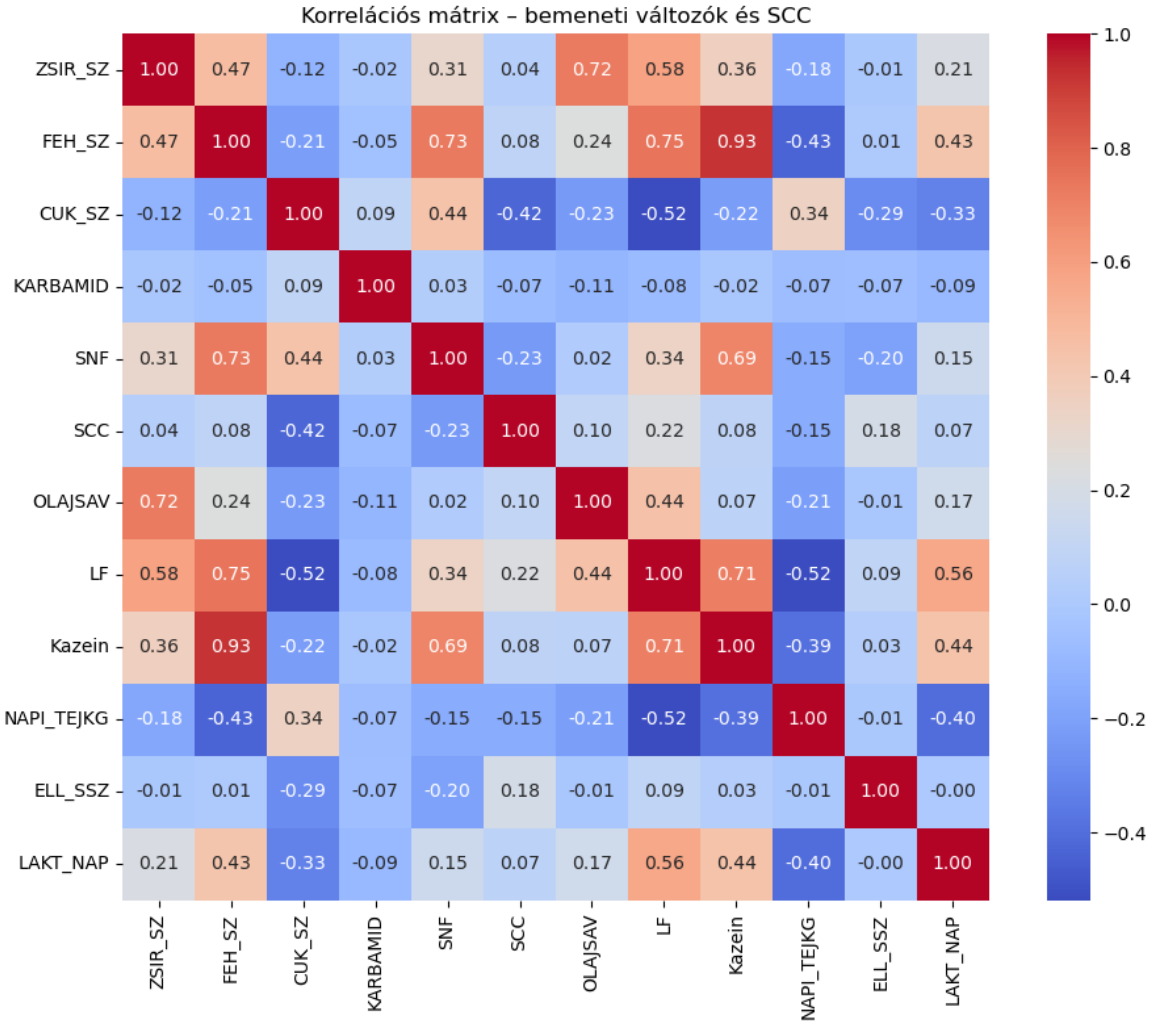
Az előzetes várakozásnak megfelelően, meglehetősen kiegyensúlyozatlan osztályeloszlású adathalmazzal találkozunk. Ebben az esetben arról van szó, hogy a

termelésben tartott tehenek nagy része egészséges, ezért a laboratóriumi minták között nagyon kevés fertőzött tejet fogunk találni. Ahhoz, hogy megfelelő modellt tudjunk készíteni, mindenképpen ki kell egyensúlyozunk az adathalmazunkat, máskülönben nagyon elfogult, pontatlan modellt kapunk.
Az adatok kiegyensúlyozására a hibrid upsampling módszert alkalmaztam a bemeneti változók varianciájának növelésével együtt.

Ezek után ki kellett választani, melyik algoritmus a legalkalmasabb a modell kialakításához. Mivel a kimeneti változónk három értéket vehet fel, olyan többosztályos (multinomiális) osztályozó algoritmusokat kellett választani, amelyek az adatokat három vagy több kategóriába sorolják. Szakirodalmi adatok alapján (*Muhammad et al.,* 2018) négyféle algoritmust hasonlítottam össze: *Random Forest, Support Vector, Decision Tree Regressor* és *Extra Trees Classifier*.
A modell tanításához az adathalmazt két részre osztottam. Az adatok 90%-a a tanító adatbázist képezte, míg 10%-át a teszteléseshez, validáláshoz használtam fel. Az adatokat véletlenszerűen osztottam két csoportra, mindkét részadatbázis eloszlása hasonló volt.

Külön-külön megvizsgáltam az egyes kimeneti osztályokban a modell teljesítményét. Végül ezen értékek átlagát tekintem a modell pontosságának. A pontosságot úgy határoztam meg, hogy a helyes találatok számát az összes kimenet számához viszonyítottam.
Ha a modell 80%-os pontosság fölött teljesít, akkor használhatónak tekintem a valós környezetben, 90%-os pontosság felett pedig akár a laboratóriumi mérések kiváltásra is alkalmas lehet.

A modell bemeneti változóinak kiválasztása során első lépésként elkészítettem a bemeneti változók közötti korrelációs mátrixot, amely segített feltérképezni az esetleges erős együttmozgásokat a változók között. Ennek célja az volt, hogy az erősen korreláló bemeneti tényezők esetleges redundanciáját felismerjem, és a későbbi modellezés során minimalizáljam az információs átfedéseket. Az eredmény az *5.6. ábrán* látható.



5.6. ábra: Az egyes változók korrelációja
a kimeneti változóhoz képest

A Python környezetben készült korrelációs mátrixot szintén a program segítségével, egy hő-térképen ábrázoltam. Ez a módszer színekkel is jelzi a választott (kimeneti) változó korrelációját a többi változóval. Így akár szemmel is láthatóvá válik az egyes változók közötti korreláció. A sötétebb színek az erősebb korrelációt jelzik.

A legjobb eredményt adó modell a következő bemeneti változókat használta: LNPC, karbamid, cukor, laktoferrin, zsír és fehérje. Az egyes modellek átlagos pontosságát az *5.8. táblázat* tartalmazza.

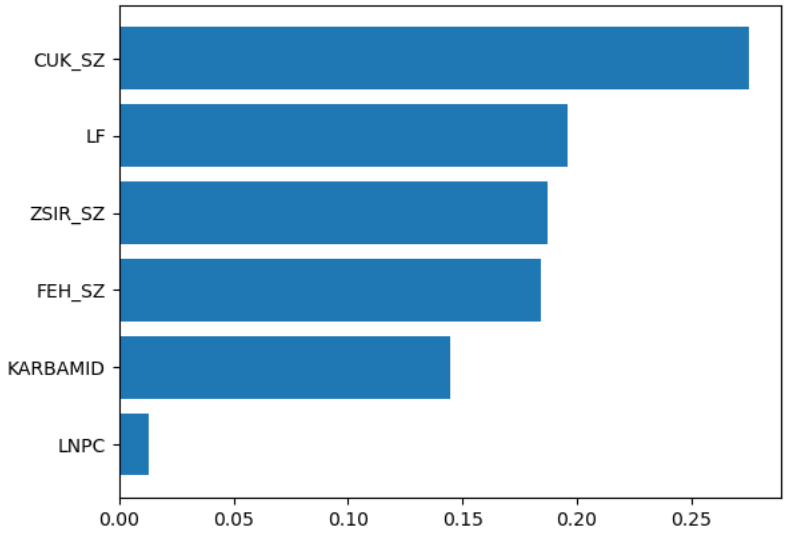
5.8. táblázat: A különböző algoritmusok által készített modellek pontosságának összehasonlítása

ML-algoritmus	**LSCC=0**	**LSCC=1**	**LSCC=2**	**Átlag**	**Recall**
Random Forest	0.88	0.86	0.85	0.86	0.75
SVM	0.86	0.85	0.80	0.84	0.72
Decision Tree Regressor	0.83	0.80	0.82	0.82	0.57
Extra Trees Classifier	0.89	0.88	0.86	0.88	0.72

Láthatjuk, hogy mindegyik kiválasztott algoritmus 80% felett teljesített. A legjobb eredményt az *Extra Trees Classifier* algoritmus segítségével készített modell adta. Fa alapú algoritmust választottam, mert ezek használhatók kategorikus változókra is, nem érzékenyek a normál eloszlásra, ezért kevesebb előzetes adat átalakításra van szükség a használatukhoz.

Fontos lenne látni, hogy a végső eredmény tekintetében melyik összetevő (melyik bemeneti változó) játssza a legnagyobb szerepet. Mikrobiológiai szempontból erre nem biztos, hogy ugyanazt a választ kapjuk, mint amit a modell végül talál. Ezért további elemzésnek vetettem alá a legjobb eredményt adó modellt.

A Python *Scikit-learn* moduljának segítségével készítettem a modell vizsgálatára egy jellemző kiemelés elemzést (feature extraction). Ez az elemző eszköz éppen arra való, hogy a végső modell bemeneti változóinak a fontosságát, a modellre gyakorolt hatását elemezze. Ebben az elemző módszerben minden bemeneti változó kap egy pontszámot, ami a fontosságát, a modell eredményére gyakorolt hatását állítja sorrendbe. Ezek alapján minden változóra kiszámoltam ezt a pontszámot és az eredményt a *5.7. ábrán* látható grafikonon ábrázoltam.



5.7. ábra: A jellemző kiemelés eredménye

A grafikonon jól látható, hogy a szomatikus sejtszám becsléséhez a *cukor*, a *zsír*, a *fehérje* és a *laktoferrin* a legmeghatározóbb bemeneti változók.

Azért volt fontos minden kimeneti lehetőségre (minden kategóriára) külön is megvizsgálni a pontosságot, hogy kiküszöböljem a modell elfogultságát (bias), tehát ne fordulhasson elő, hogy a modell csak az egészséges tejminták esetében ér el kellő pontosságot, hiszen akkor éppen a legfontosabb tulajdonsága, a ritkább - de fontosabb - fertőzött tej kimutatásának pontossága sérülne. Mivel a modell minden kategóriában hasonló pontosságot ért el, kijelenthető, hogy alkalmas lehet valós felhasználásra, hiszen kellően pontosan találja meg a fertőzött tejmintákat is.

Kutatásom igazolta, hogy készíthető olyan, gépi tanuláson alapuló modell, ami megfelelően jelzi előre a szomatikus sejtszám emelkedését. A gépi tanuláson alapuló modellt megfelelő rendszerbe illesztve folyamatosan frissítheti magát az új adatok beérkezésével, ezáltal folyamatosan javul a pontossága. Az általam kifejlesztett rendszer, amely az adattisztítástól a modellépítésen át a validálásig mindent egy rendszerben old meg, jó alapja lehet egy nagyobb méretű számítógépes rendszer kiépítésének. A feltanított modell használatához nincs szükség drága és erős hardverre, ezért akár kisebb gazdaságok is tudják alkalmazni. Ennek segítségével gyorsabbak, olcsóbbak is lehetnek a laboratóriumi vizsgálatok, és hamarabb juthatnak a tenyésztők a számukra kritikus eredményekhez, javítva ezzel tenyészetük hatékonyságát.

5.3. A kazein mennyiségének becslése

A tejtermelő gazdáknak a tej minősége több szempontból is fontos: az árképzés, az állatok általános egészsége, illetve az állatok gyógyszerezésének megtervezése (*Kunes e. al*., 2021). A tej összetevői gyorsan változhatnak a környezeti hatásoknak köszönhetően, a tehén sajátosságai alapján vagy egészségügyi rendellenességek miatt. Például a szomatikus sejtszám magasabb értékei a tejben alacsonyabb fehérjetartalmat eredményeznek (*Bisutti et al.,* 2022). Mivel a tejminőségnek fontos élelmiszer-egészségügyi jelentősége is van, a tejmintákat rendszeresen gyűjtik és akkreditált laboratóriumokban elemzik (*Kocsi et al.,* 2020). Ezért rendelkezésre áll szabványos laboratóriumi mérésekkel létrehozott, sok évre vonatkozó tejadatbázis. Magyarországon a hivatalos laboratóriumok általában a tej alábbi paramétereit mérik: teljes mikrobiális szám, szomatikus sejtszám, zsír, fehérje, fagyáspont, pH, laktóz. A tej a minőségi zsírok, fehérjék, szénhidrátok, ásványi anyagok és vitaminok gazdag forrása.

Kémiai szempontból a tej egy kolloid rendszer. Legnagyobb részben (a tehéntej kb. 90%-ban) vizet, ezenkívül fehérjéket, cukrokat és zsírokat tartalmaz. A tejet alkotó anyagok vizes oldatot képeznek, vagy a zsírcseppekben oldva vizes kolloid rendszer formájában észlelhetők.

A tejnek ezen fő alkotóelemei fontos indikátorai az állat egészségi, valamint a tej minőségi állapotának, ami felhasználható az állatgondozás javítására, a jószág takarmányozására (*Borecki et al.,* 2009).

A tejüzemekben egyre elterjedtebb a tej alkotóelemeinek a zsír és a fehérjék egységein alapuló árképzése (*Buccola et al.,* 1997). A valós idejű minőség-ellenőrzés segíti a rendellenességek korai felismerését. Egy olyan rendszer, amely a jövőre vonatkozó minőség-előrejelzést tenne lehetővé, tovább növelné a tenyésztők termelési eredményeit.

A kutatásom ezen részének a célja, hogy bebizonyítsam, a tejmintákban található kazein mennyisége előre jelezhető olyan, gépi tanuláson alapuló modell segítségével, amely a tej egyéb alkotóelemeit használja bemeneti változóként.

A tanulmányban felhasznált adatok egy magyarországi, akkreditált laboratóriumból származnak. Az adatok három különböző gazdálkodásból, hároméves periódus alatt havonta gyűjtött tejparaméterek értékeit tartalmazzák.

A függő változó (kazein) folyamatos volt, és lineáris összefüggést mutatott a legtöbb bemeneti változóval *(5.8. ábra)*, ezért egy regresszióalapú modell használata mellett döntöttem.

A kellően pontos modell feltanításához az első lépés a bemeneti változók gondos kiválasztása. A lehető legkevesebb paraméterrel érdemes dolgozni, hogy a tanítást hatékonyabbá és egyszerűbbé tegyük. Minden kiválasztott bemeneti változót átvizsgáltam a tanuló és tesztelő adatbázisok létrehozása előtt. Két bemeneti paramétert (laktációs napok és ellések száma) biológiai okokból kategorikus változóvá alakítottam. Ezek általában pontosabb modellt eredményeznek. A laktációs napokat három kategóriába osztottam (0: $n < 100$, 1: $100 < n < 200$, 2: $n > 200$). Az ellések számát szintén három kategóriába soroltam: 1, 2 és 3+.

Első lépésként adattisztítást végeztem (hiányzó adatok, kiugró értékek). A változók normalizálására nem volt szükség, mivel azokat a változókat, ahol sok kiugró érték volt, kategorikus változókká alakítottam. A kimeneti változó (kazein) folyamatos volt. A kimeneti változók legfontosabb statisztikai jellemzőit az *5.9. táblázat* tartalmazza.

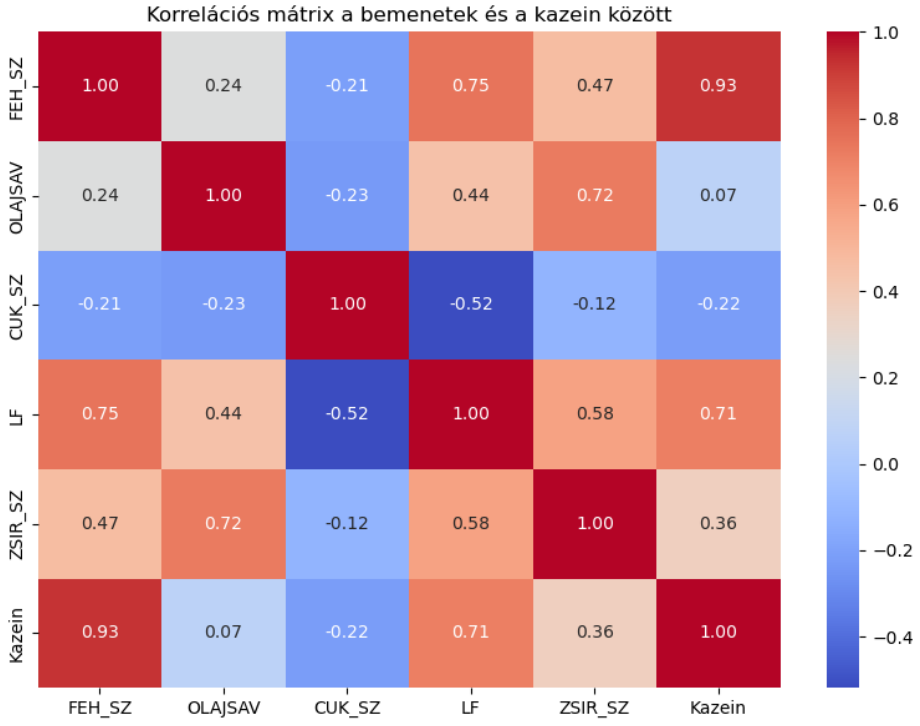
5.9. táblázat: A tejminták vizsgált paramétereinek statisztikai elemzése

	átlag	**szórás**	**min**	**25%**	**50%**	**75%**	**max**
Zsír	3,94	0,81	0	3,47	3,90	4,36	9,95
Fehérje	3,47	0,37	0	3,23	3,44	3,68	9,35
Cukor	4,95	0,30	0	4,86	5	5,12	5,69
Karbamid	22,34	7	0	18	21	25	69
SNF (zsírmentes szárazanyag)	9,07	0,43	4,97	8,84	9,07	9,32	12,04
FPD	0,45	0,12	0,09	0,30	0,50	0,56	0,67
SCC	401,70	898,79	2	47	137	343	9999
Olaj	0,74	0,23	0	0,61	0,70	0,82	4
Laktoferrin	170,10	47,47	20	140	165	194	908
Kazein	2,68	0,30	1,11	2,49	2,65	2,84	5,49
Ellések	2,17	1,35	1	1	2	3	10
Laktáció	211,81	153,79	2	93	191	294	1289

A szoftver *Pandas és Scikit-learn* használatával Pythonban íródott. A Scikit-learnből a *Lineáris Regressziós* modellt használtam a modell felépítésére. A matematikai modellre a Python széles körben elfogadott *statsmodels* modulját használtam. A tanuló

és tesztelő adatbázisokat véletlenszerűen választottam szét (90% a tanuló adatbázis és 10% a tesztelés).
Az eredmények ellenőrzésére az R^2 (*Chicco et al.*, 2021) determinációs együtthatót és az átlagos négyzetes gyökeltérést (RMSEP) használtam.

A modell elkészítéséhez meg kellett határozni a bemeneti változókat – ezek szelekciója volt a gépi tanulási projekt hatékonyságának a kulcsa. A laboratórium 14 mért tejparamétere közül kellett kiválasztani azokat, amelyek meghatározzák majd a kimeneti változót (a kazeint).
A választás megkönnyítése érdekében készítettem egy hőtérképet minden változó korrelációs értékéről *(5.8. ábra)*. A hőtérképen a magasabb értékeket (változók magasabb korrelációja) sötétebb, az alacsonyabb értékeket világosabb színekkel jelöltem.



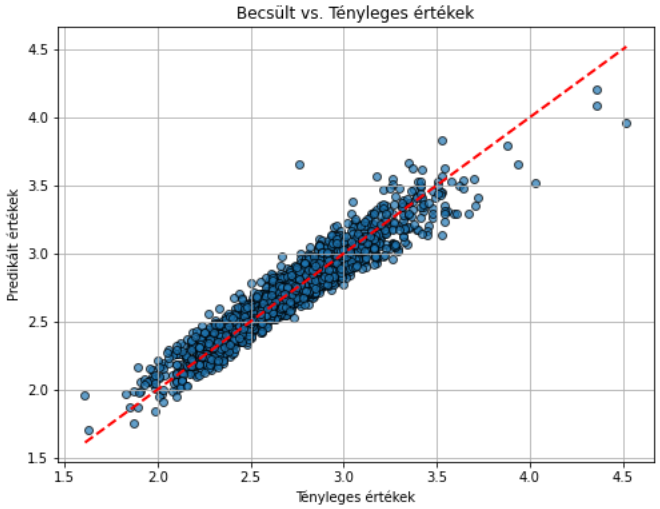
5.8. ábra: A változók hőtérképe

A hőtérkép alapján háromféle csoportot állítottam össze a bemeneti változókból, és ezek közül mindegyikre lefuttattam a modell betanítását. 10 sorozatot (tanítás–ellenőrzés) futtattam minden egyes bemeneti csoporttal úgy, hogy az adatbázist minden alkalommal véletlenszerűen osztottam fel tanító és ellenőrző részekre. Az *5.10. táblázat* a futások legjobb és legrosszabb eredményeit tartalmazza.

5.10. táblázat: A különböző bemeneteken alapuló előre jelzéseink eredményei

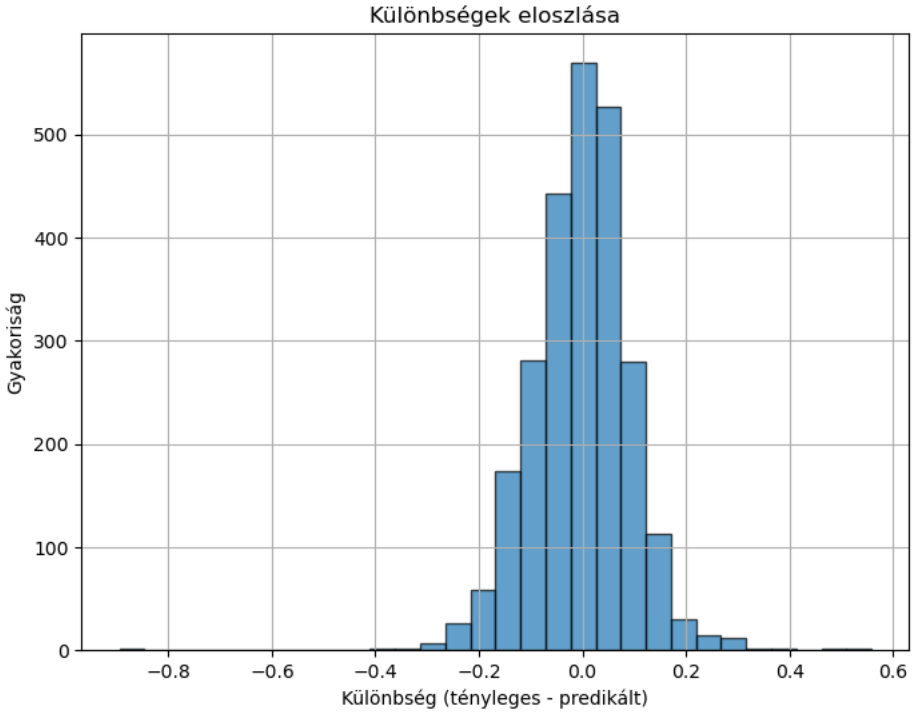
Bemeneti változók	**R^2**	**MRSE**
'Laktációs napok', 'Karbamid', 'Cukor', 'LF', 'Zsír'	0.82/0.78	0.033
' Feherje', 'Olaj', 'Cukor', 'LF', 'Zsír'	0.86/0.84	0.018
'Laktációs napok', 'SNF', 'LF', 'Zsír'	0.83/0.76	0.034

A valós és a modell által számított kimeneti értékeket egy közös grafikonon is ábrázoltam *(5.9. ábra)*. A piros vonal jelzi az egyes kimeneti értékekből képzett egyenest, a kék pöttyök pedig a becsült értékeket.



5.9. ábra: A (mért) célértékek és a becsült értékek a modellünk alapján

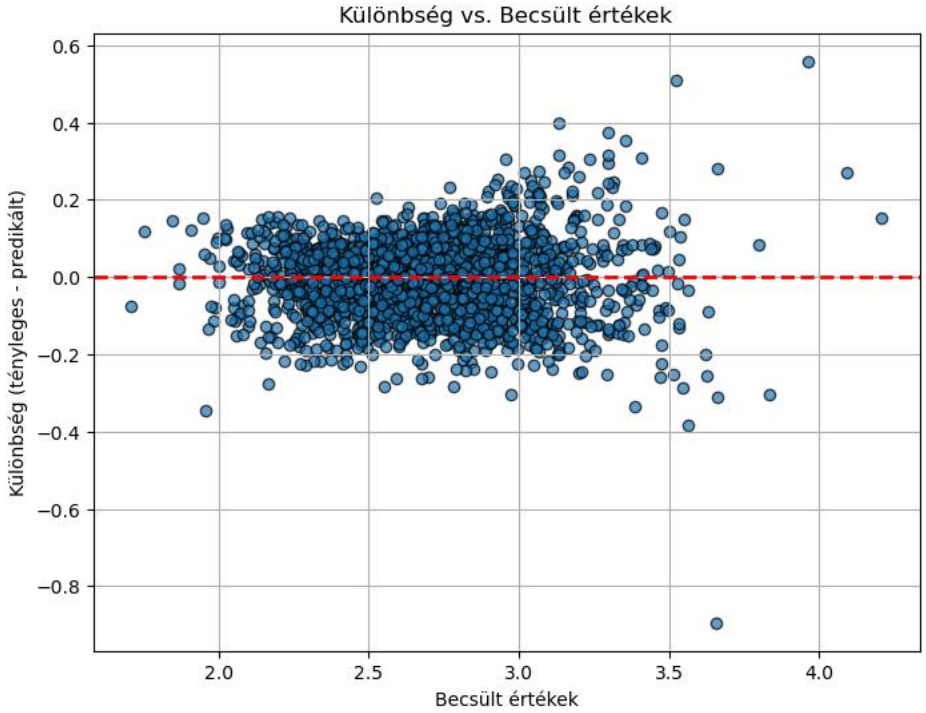
A becsült és tényleges értékek különbségének az eloszlása az *5.10. ábrán* látható.



5.10. ábra: A becsült és tényleges értékek különbségének eloszlása

Azt is megvizsgáltam, hogy a modell mennyire jól ragadja meg a változók közötti kapcsolatot. Erre a célra a különbségek vs. becsült értékek diagramját használtam, ami a regressziós modell validációjának egyik fontos eszköze. Az *5.11. ábrán* látható grafikonon az *x* tengelyen a modell által előre jelzett (becsült) értékek, az *y* tengelyen pedig a különbségek, azaz a tényleges értékek és a modell által becsült értékek különbségei (hibák) szerepelnek.

Láthatjuk, hogy a pontok eloszlása véletlenszerű a 0 körül, vagyis a modell képes jól megragadni a változók közötti kapcsolatot. A hibák normális eloszlást mutatnak, így kijelenthetjük, hogy megfelelően specifikált modellről van szó.



5.11. ábra: Különbség vs. becsült értékek grafikonja

Az egyes bemeneti változók fontosságának meghatározásához továbbfejlesztettem a programot úgy, hogy kiírja a végső regressziós egyenlet együtthatóit is. Így pontos ismeretünk van a modell által kialakított regressziós modellről.

$$y = 0.67x_1 - 0.315x_2 + (-0.0058)x_3 + 0.00119x_4 + 0.016x_5 + 0.38 \qquad (5.1)$$

ahol x_1 = fehérje, x_2 = olajsav, x_3 = cukor, x_4 = laktoferrin, x_5 = zsír.

Az adatbázis 25 000 mintát tartalmazott, ezt használtam a modell feltanításához a legjobb eredmény elérése érdekében. Ha több adatot tudunk használni, esetleg más laboratóriumokból származókat is, akkor a modell ugyanezzel a rendszerrel tovább pontosítható, általános teljesítménye tovább javítható, hiszen a feldolgozás teljesen automatizált. A bemeneti változók többféle összeállítását is megvizsgáltam az optimális modell megtalálásáért. Több adattal tanítva modellünket, a precizitás növelhető. Az eredmények azt mutatják, hogy az 'olaj', a 'cukor', a 'laktoferrin' és a 'zsír' használatával olyan modellt építhetünk, ahol $R^2 = 0.86$, és ehhez kis hibaérték is tartozik. A gépi tanuláson alapuló modell a kazein mennyiségének becslésére jobb eredményeket hozhat, mint a hagyományos módszerek. A kazein mennyiségének meghatározására általában a Fourier-transzformációs infravörös (FTIR-)

spektroszkópiát használják. Ennél a módszernél az eredményesség irodalmi adatok alapján $R^2 = 0.73$ (*Sørensen et al.,* 2003), vagyis a gépi tanulással feltanított modell ennél jobb eredményt ér el.

A predikciós eljárást összehasonlítottam egy hagyományos matematikai becsléssel is. Ezt a matematikai megoldást alkalmazva az eredmény $R^2 = 0.8395$.

A gépi tanuláson alapuló előrejelzések 86%-os pontossága jobb eredményt ad, mint az általam vizsgált matematikai módszer 83,95%-os eredménye. Ez az eredmény általános használatra (pl. a tej átvételénél árazásra) megfelelő, amennyiben a pontos kazeinmennyiséget nem szükséges meghatározni. A kutatás során az általam fejlesztett megoldás automatizálja az adattisztítás teljes folyamatát, és az adatátalakítást és előrejelzést egyetlen rendszer keretében megoldja.

A modellt nagyon egyszerűen lehet tovább hangolni, tanítani új adatokkal, csupán a bemeneti táblázatot kell a megadott formátumra hozni. A regressziós modell fejlesztésének eredményeit később további kutatások esetében is jól tudom majd hasznosítani. Ha egy nagyobb mintán, több komplexebb jellemzővel bíró adatbázis alapján tervezünk modellt fejleszteni akkor ezek az eredmények jó kiindulási alapot szolgáltatnak pl. egy neurális háló alapú modell kifejlesztéséhez. Annak eredményeit tudjuk majd ehhez hasonlítani, illetve jó képet kaphatunk a lehetséges bemeneti változók köréről.

6. ÚJ TUDOMÁNYOS EREDMÉNYEK

Kutatómunkám során számos összefüggést tártam fel a meglévő szarvasmarha-tenyésztés során keletkezett adatbázisok hasznosítására. A mesterséges intelligencia és ezen belül a gépi tanulás kiválóan alkalmas olyan modellek feltanítására, amelyek segítségével a tenyésztők, termelők számukra fontos tenyésztési, minőségi vagy egészségügyi paraméterek becslésére és előrejelzésére lesznek képesek, akár egy egyszerű alkalmazásban.

6.1. Tézis

1. Küllemi bírálati pontok előrejelzése

Kutatásom során létrehoztam és validáltam egy modellt, amely bizonyítja, hogy egy gépi tanulásos regressziós modell képes objektív módon és magas pontossággal reprodukálni a hagyományos, szakértői küllemi bírálati értékeléseket.

A szakértői pontozás szubjektív lehet, amit befolyásolhatnak az egyéni tapasztalatok, vélemények vagy akár eltérő módszertan is. A kutatásom eredménye bizonyítja, hogy a gépi tanulási megközelítés képes az ilyen emberi variabilitást csökkenteni, így objektív és megismételhető eredményeket ad, ami fontos előrelépés az állattenyésztés értékelési módszereiben.

Az általam feltanított és ellenőrzött modell eredményeit az *6.1. táblázatban* láthatjuk.

6.1. táblázat: A legjobb eredményt adó modell pontossága az egyes testtájak pontszámértékeire (n = 325)

Modell típusa	**Függő változók**	**Átlagos négyzetes hiba**	**R^2**
Lineáris regresszió	Izmoltság (0–60)	3,38	0,86
	Far hossza (1–9)	0,21	0,93
	Mellkas izmoltsága (1–9)	0,35	0,86
	Far izmoltsága (1–9)	0,41	0,77
OLS modell	Far hossza	0,33	0,91

A modell építéséhez használt adatbázis jellemzése:

Adatforrás: Limousin és Blonde d'Aquitaine Tenyésztők Egyesülete (Budapest), anonimiziált adatbázis

Állatok az adatbázisban: 18 bika utódai, amelyek 1990 és 1996 között születtek

Összes mérési eredmény száma: 325

Adatbázis adatai: Az adatbázisban szereplő változók leírását a *6.2. táblázat* tartalmazza.

6.2. táblázat: Az adatbázisban szereplő változók leírása

Változó megnevezése az adatbázisban	Jelentése
H_MARMAG	Marmagasság
H_MELLMELY	Mellkasmélyég
H_VALLFESZ	Vállfeszesség
H_VALLKOT	Vállkötés
H_LABSZERK	Lábszerkezet
H_CSONT	Csontfinomság
Hasznalati_osszes	Használati értékek összpontszáma
HO_TEST	Testhosszúság
HO_HAT	Háthosszúság
HO_AGYEK	Ágyékhosszúság
HO_FAR,	Farhosszúság
hossz ossz	Hosszúsági értékek összpontszáma
SZ_MAR	Marszélesség
SZ_MELL	Mellkasszélesség
SZ_AGYEK	Ágyékszélesség
SZ_FAR1	Farszélesség I.
SZ_FAR2	Farszélesség II.
SZ_FAR3	Farszélesség III.
szel_ossz	Szélességi értékek összpontszáma
IZ_SZUGY	Szügyizmoltság
IZ_LAPOCKA	Lapocka izmoltsága
IZ_HAT	Hátizmoltság
IZ_FAR	Farizmoltság
IZ_COMBH,	Combhosszúság
IZ_COMBT	Combteltség
izom_ossz	Izmoltsági értékek összpontszáma
Kullemi_biraalati_osszpontszam	Küllemi bírálat összpontszáma

A legjobb eredményt adó modell bemutatása:

Felhasznált Python-modul: Scikit-learn

Felhasznált algoritmus: LinearRegression()

test_size = 0.10: Teszteléshez használt adatok aránya (10%)

random_state = 10: Ez a paraméter biztosítja a véletlenszerű felosztás reprodukálhatóságát, azzal, hogy a véletlenszám-generátor magját (seed) mindig ugyanarra a kiindulási értékre állítja.

Intercept (konstans tag): -0.079789

Koefficiensek (súlyok): [0.22136146; -0.56044067; -0.66474507]

Az automatizált modell alkalmazása gyorsabb és könnyen skálázható megoldást kínál a küllemi értékelések elvégzésére. Ez lehetőséget teremt a nagyobb adatmennyiséggel rendelkező állatállományok objektív összehasonlítására, és hatékonyabb döntéstámogatást nyújt a tenyésztési programok kialakítására.

A kutatásom innovatív módon integrálja a hagyományos szakértői adatbázist a modern gépi tanulási technikákkal. A modell teljesítményének és a bemeneti változók szerepének elemzése révén azonosíthatók azok a küllemi paraméterek, amelyek a legnagyobb hatással vannak az értékelésre. Ez új tudományos ismereteket nyújt arról, hogy mely jellemzők lehetnek a legfontosabbak a szarvasmarhák genetikai és gazdasági értékének meghatározásában.

A modell sikeressége azt bizonyítja, hogy a gépi tanulás nemcsak kiegészítheti, de helyettesítheti is a hagyományos, szakértőalapú értékelést, ezzel hozzájárulva egy modern, adatvezérelt állattenyésztési értékelési rendszer kialakításához.

6.2. Tézis

2. A szomatikus sejtszám előrejelzése

Kutatásom során bizonyítottam, hogy lehetséges olyan, gépi tanuláson alapuló modellt kifejleszteni, ami képes megbízhatóan eldönteni a tejminták alapján, hogy a tehén egészséges vagy fertőzött, anélkül hogy a szomatikus sejtszám közvetlen mérésére szükség lenne.

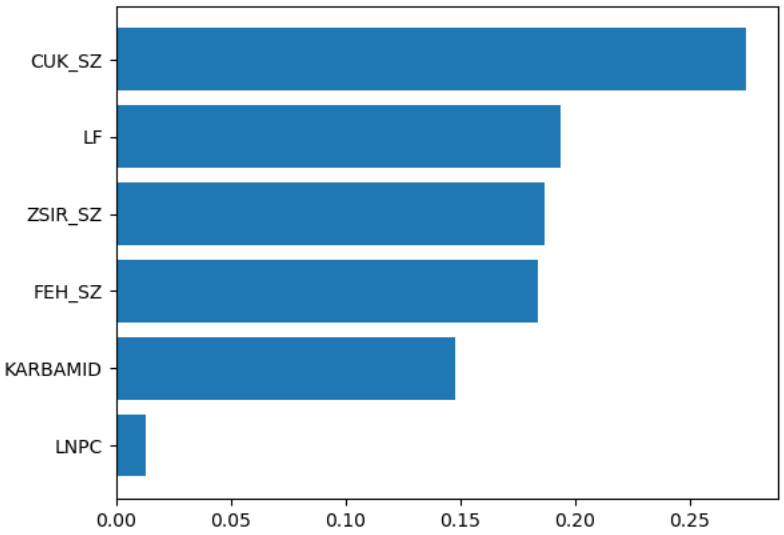
A kutatásom során egy gépi tanuláson alapuló modellt hoztam létre 25 000 tejminta adatai alapján, ami más, mérhető tejparamétereket használva előre jelzi a szomatikus sejtszám mértékét.

A kutatás során bebizonyítottam, hogy más tejparaméterek segítségével javíthatom a masztitisz korai felismerésének lehetőségét, miközben a gépi tanulás alkalmazása csökkenti a laboratóriumi módszerek költségeit és időigényét.
Az általam vizsgált algoritmusok közül az Extra Trees Classifier algoritmussal tanított modell hozta a legjobb eredményeket, amelynek eredményeit az *6.3. táblázatban* láthatjuk.

6.3. táblázat: R^2 eredmények a szomatikus sejtszám osztályai szerint, illetve az átlagos R^2 érték

ML-algoritmus	**LSCC=0**	**LSCC=1**	**LSCC=2**	**Átlag**
Extra Trees Classifier	0.89	0.88	0.86	0.88

Azt is megvizsgáltam, hogy a tanításhoz igénybe vett paraméterek közül melyiknek mekkora hatása volt a végső modellre. Az eredményt az *6.1. ábrán* láthatjuk.



6.1. ábra: A tejminták osztályozása szomatikus sejtszám alapján, a modell kialakításában legfontosabb szerepet játszó bemeneti változók fontossága

Az eredmények egyértelműen mutatják, hogy a tej egyéb paramétereiben rejlő információ elegendő ahhoz, hogy a gépi tanulási modell megbízhatóan osztályozza a mintákat „egészséges” vagy „fertőzött/masztitiszes” kategóriába.

A modell építéséhez használt adatbázis jellemzése:

Adatot szolgáltató tenyésztők száma: 3

Évek: 2019–2021

Rendszeresség: Havi egyszeri tejminőség-elemzés (tenyésztőnként 36 havi adat)

Összes mérési eredmény száma: 25 000

Mért paraméterek száma: 14

A modell részletes jellemzése:

n_estimators = 100 (Az *n* azt határozza meg, hogy hány döntési fát épít a modell az együttes tanuláshoz. Minél több fa épül, annál robusztusabbá válik a modell, mert az egyes fák hibái részben kioltódnak, de ezzel arányosan nő a számítási igény is.)

test_size = 0.10: Teszteléshez használt adatok aránya (10%)

random_state = 10

Feature importances (bemeneti változók súlya): [0.01788029; 0.10898641; 0.13511609; 0.24754355; 0.1710435; 0.15843071]

Felhasznált jellemzők száma: 6

Cél osztályok: [0 1 2]

Maximális mélységek az egyes fákban: [48, 48, 41, 44, 46, 43, 47, 39, 51, 40, 43, 40, 41, 37, 43, 41, 43, 43, 40, 45, 42, 46, 46, 42, 42, 44, 42, 41, 41, 44, 47, 46, 45, 42, 37, 44, 44, 40, 41, 44, 39, 42, 42, 43, 44, 41, 39, 45, 46, 47, 49, 44, 40, 41, 36, 40, 46, 48, 38, 45, 43, 44, 46, 44, 37, 41, 47, 36, 43, 40, 42, 48, 41, 40, 45, 38, 41, 46, 46, 40, 44, 42, 38, 41, 40, 42, 39, 38, 44, 43, 44, 39, 43, 46, 47, 38, 40, 39, 45, 45]

6.3. Tézis

3. Kazein mennyiségének előrejelzése

A kutatásom igazolta, hogy egy regressziós, gépi tanuláson alapuló modell segítségével – kizárólag a tej egyéb, rutinszerűen mért paramétereiből – megbízhatóan előre jelezhető a kazein mennyisége. Nem szükséges közvetlenül, költségesebb módon meghatározni a kazeintartalmat, hanem az indirekt mérési adatokból is kinyerhető a mennyiségre vonatkozó információ.

A modell segítségével nemcsak a kazein mennyiségének előrejelzése vált lehetővé, hanem meghatároztam azt is, hogy mely laboratóriumi mérések játszanak szerepet a kazeintartalom predikciójában. Ez a feature importance elemzés új ismereteket szolgáltat arról, hogy a tejminőség szempontjából mely jellemzők a legkritikusabbak.

6.4. táblázat: A legjobb eredményt adó bemenetek a kazeintartalom becsléséhez

Bemeneti változók	**R^2**	**MRSE**
'Feh', 'Olaj', 'Cukor', 'LF', 'Zsír'	0.86/0.84	0.018

A kazeintartalom pontos előrejelzése kulcsfontosságú a sajtgyártásban és a tej minőségalapú árképzésénél. Az automatizált, adatvezérelt módszerrel nemcsak a minőség-ellenőrzést lehet gyorsítani és objektívebbé tenni, hanem potenciálisan költséghatékonyabbá is válik az ipari folyamatok ellenőrzése. Ez az eredmény különösen releváns a tejiparban, ahol a termékminőség meghatározó tényező a piaci érték és a fogyasztói elégedettség szempontjából.

Az egyes bemeneti változók fontosságának meghatározásához továbbfejlesztettem a programot úgy, hogy kiírja a végső regressziós egyenlet együtthatóit is. Így pontos ismeretünk van a modell által kialakított regressziós modellről.

$$y = 0.67x_1 - 0.315x_2 + (-0.0058)x_3 + 0.00119x_4 + 0.016x_5 + 0.38 \qquad (6.1)$$

ahol x_1 = fehérje, x_2 = olajsav, x_3 = cukor, x_4 = laktoferrin, x_5 = zsír.

A modell építéséhez használt adatbázis jellemzése:
Adatot szolgáltató tenyésztők száma: 3
Évek: 2019–2021
Rendszeresség: Havi egyszeri tejminőségelemzés (tenyésztőnként 36 havi adat)
Összes mérési eredmény száma: 25 000
Mért paraméterek száma: 14

A legjobb eredményt adó modell bemutatása:
Felhasznált Python-modul: Scikit-learn
Felhasznált algoritmus: LinearRegression()
test_size = 0.10: Teszteléshez használt adatok aránya (10%)
random_state = 10: Ez a paraméter biztosítja a véletlenszerű felosztás reprodukálhatóságát, azzal, hogy a véletlenszám-generátor magját (seed) mindig ugyanarra a kiindulási értékre állítja.

Intercept (konstans tag): 0.38

Koefficiensek (súlyok): [6.71812578e-01 -3.15429010e-01 -5.58654963e-03 1.19884915e-03 1.60130038e-02]

6.4. Tézis

4. Osztályeloszlás kiegyenlítése valós tejminőségi adatokon

A modellek megalkotásához használt adatbázisok erősen kiegyensúlyozatlanok voltak. Kutatásom során bebizonyítottam, hogy hibrid adatkiegyensúlyozási technikákkal ezek az „elfogult adatbázisok” is transzformálhatók úgy, hogy alkalmasak legyenek gépi tanulási algoritmusok feltanításához.

A kutatás során alkalmazott hibrid megközelítés – amely ötvözi az új, releváns adatok bevonását, mesterséges adatok injektálását (adataugmentáció) és a bemeneti változók finomhangolását is – új megoldást kínál az állategészségügyi szempontból kiegyensúlyozatlan, tejminőséget vizsgáló adatbázisok kezelésére. Ezzel a módszerrel sikerült csökkenteni a túlzottan reprezentált „egészséges” osztály hatását, és kiegyensúlyozottabb, reprezentatívabb adathalmazt létrehozni.

A kutatásom rámutatott, hogy az adatbázis kiegyensúlyozatlansága (azaz a nagyon alacsony arányú fertőzött minták) hogyan torzíthatja a prediktív modellek tanulási folyamatát, és milyen módszerekkel lehet ezt a problémát kezelni.

7. KÖVETKEZTETÉSEK ÉS JAVASLATOK

Kutatási eredményeim hozzájárulnak a tejminőség, az állategészség és a tenyészeti értékelés adatvezérelt megközelítésének további fejlesztéséhez.

A gépi tanulásos modellek – akár a küllemi értékelés, akár a masztitisz korai diagnosztikája, vagy a kazeintartalom előrejelzése terén – képesek objektív és megismételhető eredményeket produkálni, amelyek csökkentik az emberi szubjektivitásból adódó hibákat. A gépi tanuláson alapuló modellek azonos vagy jobb eredményt produkálnak, mint a hagyományos matematikai modellek, ezt kutatásom is alátámasztja. A mesterséges intelligencián alapuló modellek azonban sokkal jobban fognak teljesíteni nagyméretű adatbázisok esetén (gyorsabban taníthatóak), valamint az újabb adatok érkezésével az modell finomhangolása, fejlesztése folyamatosan elvégezhető. Ezért a kutatásom bizonyította, hogy további adatok gyűjtése esetén, esetleg valóban nagyméretű adatbázisok kialakítása után iparilag is jól hasznosítható modellek (és ezek alapján applikációk) készíthetők a dolgozatomban leírt módszerrel. További kutatásokban érdemes lehet más, eddig kevésbé vizsgált tényezőket is bevonni (például genetikai, környezeti paramétereket), hogy a modellek még átfogóbb képet kapjanak az állatok állapotáról.

Új adatok, nagyobb adathalmazok rendelkezésre állása esetén további gépi tanulási algoritmusok (pl. ensemble módszerek, mélytanulás) összehasonlítása is célszerű lenne még komplexebb, általánosabban használható modellek kialakításához. Az általam fejlesztett szoftver környezet, moduláris felépítése miatt agyszerű átalakítások után alkalmas másmilyen algoritmusok használatára is, illetve könnyedén kapcsolható nagyobb méretű adatbázisokhoz is.

A küllemi bírálati pontozáshoz használt módszertan nemcsak az állatok küllemi pontszámának becslésében lehet hasznos, hanem alkalmazható más hasonló, emberi értékeléstől függő területeken is (például más állatfajták, illetve fajok értékelésében, minőség-ellenőrzésben stb.).

Az adatvezérelt döntéstámogatás az állattenyésztésben és a tejiparban egyszerűen megvalósítható, illetve továbbfejleszthető, ezt a modelljeim pontossága is jól jelzi. Mivel mind az állattenyésztés, mind a tejipar hatalmas historikus adatvagyonnal rendelkezik, ezek rendszerezésével, strukturált tárolásával további modellek készíthetők.

Az általam kifejlesztett kísérleti szoftverkörnyezet – mely az adatfeldolgozástól, a modelltanításon át az elemzésig és validálásig, mindent egy környezetben old meg, más elemző rendszerek bemeneteként is szolgálhat. A kimeneti eredmények például összekapcsolhatók gazdasági elemzésekkel a tej minőségalapú árképzéséhez, hogy az iparág számára még konkrétabb hasznot biztosítsanak. A modellek tanítása az általam kialakított szoftverkörnyezet továbbfejlesztésével lényegesen nagyobb mennyiségű adat esetén is megoldható. A feltanított modell használatához nincs szükség különleges hardvereszközökre, ezért az ebből kialakított alkalmazások akár egy egyszerűbb laptopon vagy táblagépen is futtathatók. Az így kialakított alkalmazások használata tehát akár kisebb gazdaságok számára is lehetségessé válik, javítva ezzel azok termelékenységét.

A hibrid adatkiegyensúlyozási megközelítésem – amely új adatok bevonásával, mesterséges minták generálásával és a bemeneti változók finomhangolásával működött – bizonyítja, hogy a kiegyensúlyozatlan adatbázisok problémája jól kezelhető a kevés beteg állat adatait tartalmazó adatbázisok esetében is, és ez javítja a diagnosztikai modellek teljesítményét. Vagyis ezek a megoldások nemcsak szintetikus adatbázisok esetében adnak jó eredményt, hanem a biológiai rendszerekből alkotott adatbázisok is feljavíthatók ilyen módszerekkel.

Az eredményeim bizonyítják, hogy a modern gépi tanulási technikák sikeresen integrálhatók az állattenyésztési és tejipari folyamatokba. A jövőben érdemes még szorosabb együttműködést kialakítani a biológusokkal, állatorvosokkal és mérnökökkel, hogy az interdiszciplináris kutatások révén továbbfejlesszük ezeket a folyamatokat.

A dolgozatomban bemutatott új tudományos eredmények nemcsak a meglévő módszerek hatékonyságát bizonyítják, hanem számos új irányt is nyitnak a jövőbeli kutatások és gyakorlati fejlesztések számára, amelyek tovább javíthatják az állatok értékelését, az egészségügyi diagnosztikát és a tejipar versenyképességét.

8. ÖSSZEGZÉS

Kutatásaim során a szarvasmarha-tenyésztésben és a tejtermelésben hasznosítható adatalapú mesterséges intelligencia modellek kialakításával és vizsgálatával foglalkoztam.

A kutatás az adatok felkutatásával és elemzésével kezdődött. Minden adatbázist át kellett alakítani, a bennük lévő adatokat egyenként megvizsgálva a gépi tanulás számára is jól használható adatbázist kellett kialakítani.

Elsőként a Limousin marhák küllemi bírálati pontszámainak becslésével foglalkoztam. Bizonyítottam, hogy készíthető olyan modell, amely bizonyos küllemi pontszámokat igen kimagasló pontossággal képes megbecsülni.

A becslési eredmények elemzése alapján elmondható, hogy egy jól feltanított modell alkalmas a szarvasmarhák küllemi bírálati pontjainak becslésre, más pontszámok ismeretében. Vagyis kevesebb méréssel, kevesebb munkával végezhető teljes körű pontozás.

További adatok gyűjtése és más farmok adatainak felhasználása is szükséges lenne az elkészült modell általánosításhoz. Érdemes lenne megvizsgálni, hogy más országok hasonló pontozási eredményeinek bevonásával milyen eredmények születnek.

A modell kutatási célú kialakítása és validálása mindazonáltal jól bizonyította, hogy hasonló MI-modellek segíthetik, támogathatják a szakértőket és tenyésztőket is az informatikailag is támogatott precíziós gazdálkodás kialakításában. Az általam kifejlesztett rendszer alapját képezheti akár egy mobilapplikációnak is, amellyel bármely helyszínen elvégezhető az egyszerűsített szakértői pontozás.

Kutatásom másik részében bizonyítottam, hogy a tejminősítő laboratóriumok begyűjtött tejadatai felhasználhatók a szomatikus sejtszám előrejelzésére. Készítettem egy olyan, gépi tanuláson alapuló modellt, amely más tejparaméterek alapján kellő pontossággal becsüli a tej szomatikussejtszám-tartalmát. Ezek alapján a rendszer jó eredménnyel sorolta a teheneket fertőzött/nem fertőzött kategóriákba. A laktotranszferrin – amely biológiai szempontból a legjelentősebb hatást gyakorolja az SCC értékekre – sokkal jobb eredményt ad, ha más bemeneti változókkal együtt használjuk. A legjobban teljesítő modell vizsgálata kimutatta, hogy a laktotranszferrin nem a legjelentősebb bemeneti változó a modell szempontjából. A modell pontossága bebizonyította, hogy a jelenleg is mért tejparaméterek alapján lehetséges az SCC

értékét előre jelezni. A tenyésztők szempontjából ez azt jelenti, hogy alkotható olyan modell, ami képes megjósolni a tőgygyulladás kialakulását, így több időt biztosít a gazdáknak arra, hogy megfigyeljék a veszélyeztetett teheneket, és így hamarabb megkezdhessék a kezelést. Ehhez azonban további, időbélyeggel ellátott adatokra lesz szükség, amelyeket a fejőrobotokat használó gazdaságokból lehet begyűjteni.

Továbbá bebizonyítottam, hogy a logaritmikus SCC érték három kategóriába történő osztása jól alkalmazható előrejelzési célokra, és egyúttal megfelel a piac tejegészségügyi követelményeinek is. Mivel a gyűjtött tejminták többsége egészséges tehenektől származik, az bemeneti adathalmaz kiegyensúlyozása elengedhetetlen a gépi tanulási algoritmus tanítása előtt. A szokásos kiegyensúlyozási módszerek sikeresnek bizonyultak az adatok egyensúlyának megteremtésében.

Eredményeim bizonyítják, hogy a többosztályos gépi tanulás hatékonyan alkalmazható a tejmintákban előforduló magas szomatikus sejtszám előrejelzésére. Ez az algoritmus alapul szolgálhat a szomatikus sejtszám jövőbeli előrejelzéséhez, és hozzájárulhat a tehenek egészségi állapotának monitorozásához. Ha a gazda már azelőtt elkezdheti a gyógyszeres kezelést, hogy a tejminták szomatikus sejtszáma kritikus értékeket érne el, az kevesebb gyógyszer felhasználását és hatékonyabb termelést tesz lehetővé.

A kutatásom harmadik részében a kazein mennyiségének előrejelzésével foglalkoztam. Bebizonyítottam, hogy meglévő tejminták adataiból készíthető olyan, gépi tanuláson alapuló modell, amely jó eredménnyel becsüli a tejben a kazein mennyiségét.

Jelenleg főleg kémiai módszereket használnak a tejben található kazein mennyiségének meghatározására. A legtöbb mérés izoelektromosan ülepíti a tejben a kazeint 4,6 pH-értéken és szűréssel választja el a kazeint a nemkazein részektől (AOAC International (2016) standard módszerek 990.20 és 990.21). A kazeintartalom számítása történhet a teljes fehérjemennyiség és nemkazein fehérjék különbségének kiszámításával a Kjeldahl-alapú módszerrel vagy az elkülönített kazeinoldat közvetlen méréséből.

A Kjeldahl-alapú módszert 1938 óta alkalmazzák ipari standard módszerként a kazein mennyiségének megállapítására. Az eljárást alapvetően a nitrogén százalékos mennyiségének meghatározására használják szerves vegyületekben. Az eljárás során a vegyületet kénsavval és réz(II)-szulfát katalizátorral felforralják, melynek hatására a nitrogén ammónium-szulfáttá alakul. Majd lúgot adnak az oldathoz, és a felszabaduló

ammóniát desztillálják. Az ammóniát ezután valamilyen savat tartalmazó oldatba vezetik, és mennyiségét a nem reagált sav mennyiségének titrálásával határozzák meg. Ebből az eredményből az eredeti anyag nitrogéntartalma meghatározható. A módszert Johan Kjeldahl (1849–1900) dán kémikus dolgozta ki.
Léteznek szeparációs technikákon és infravörös spektroszkópián alapuló eljárások a kazeinmennyiség mérésére, ezek a megoldások meglehetősen bonyolultak (*Dimenna et al.,* 1981; *Bonfatti et al.,* 2008). Az általam kidolgozott, adatalapú módszer lényegesen egyszerűbb a kazeintartalom becslésére, ezért nagyon hasznos lehet a tejiparban. A modell további fejlesztése (időbeni előrejelzéssé) pedig segítheti a tenyésztés hatékonyságát.

A megoldás eredményességéből látható, hogy a gépi tanulási módszerekkel kialakított modellek segíthetnek becsülni, előre jelezni a kémiai és biológiai értékeket a tejben. A módszer előnye, hogy már meglévő adatbázisok feldolgozásával alakítja ki a modellt, így költséghatékony módon fejleszthető. Mivel az adatok gyűjtése nem igényel új beruházást vagy a korábbitól eltérő mintavételezési technikát, ezért az általam fejlesztett rendszer könnyedén beilleszthető bármely termelési módszerbe.

9. SUMMARY

During my research, I focused on developing and analyzing data-driven artificial intelligence models applicable to cattle breeding and milk production.

The research began with data collection and analysis. Each database had to be transformed, and the data within had to be examined individually to create a machine-learning-friendly dataset. First, I worked on estimating the conformation scores of Limousin cattle. I proved that a model can be created that estimates certain conformation scores with outstanding accuracy. The analysis of the estimation results demonstrated that a well-trained model can predict the conformation scores of cattle based on known scores of other traits. This means that a complete evaluation can be performed with fewer measurements and less labor.

To generalize the developed model, further data collection and the inclusion of data from other farms are necessary. It would also be valuable to analyze similar scoring results from other countries to assess potential outcomes. The design and validation of the research-oriented model clearly demonstrated that similar AI models can assist and support experts and breeders in establishing precision farming with IT support. The system I developed could even serve as the foundation for a mobile application that enables simplified expert scoring at any location.

In another part of my research, I demonstrated that milk quality data collected by dairy laboratories can be used to predict somatic cell count (SCC). I developed a machine learning model capable of estimating the SCC content in milk with sufficient accuracy based on other milk parameters. The system successfully classified cows into infected and non-infected categories.

Lactotransferrin, which has the most significant biological impact on SCC values, produced much better results when used alongside other input variables. The examination of the best-performing model revealed that lactotransferrin is not the most critical input variable from the model’s perspective. The model's accuracy proved that SCC values can be predicted based on currently measured milk parameters.

From a breeder's perspective, this means that a model can be created to predict mastitis development well in advance, giving farmers more time to monitor at-risk cows and initiate treatment sooner. However, additional time-stamped data will be required, which can be collected from farms using milking robots.

Additionally, I proved that categorizing logarithmic SCC values into three groups is suitable for prediction purposes while also meeting the milk hygiene standards of the market. Since most collected milk samples come from healthy cows, balancing the input dataset is essential before training a machine learning algorithm. Standard balancing methods have been successfully applied to ensure data equilibrium.
My findings demonstrate that multi-class machine learning can effectively predict high SCC occurrences in milk samples. This algorithm can serve as a foundation for future SCC predictions and contribute to monitoring cow health. If farmers can begin treatment before SCC levels reach critical thresholds, it will lead to reduced drug use and more efficient production.

In the third part of my research, I focused on predicting casein content. I proved that an AI-based model can be developed using existing milk sample data to estimate the casein content in milk with high accuracy.
Currently, casein content in milk is primarily determined using chemical methods. Most measurements involve iso-electrically precipitating casein at a pH of 4.6 and separating it from non-casein components via filtration (AOAC International (2016) standard methods 990.20 and 990.21). Casein content can be calculated using the difference between total protein and non-casein proteins via the Kjeldahl-based method or through direct measurement of the isolated casein solution.
The Kjeldahl-based method has been the industrial standard for casein content determination since 1938. It is primarily used to measure nitrogen content in organic compounds. The process involves boiling the compound with sulfuric acid and a copper (II) sulfate catalyst, converting nitrogen into ammonium sulfate. Then, alkali is added, and the released ammonia is distilled. The ammonia is absorbed into an acid solution, and its quantity is determined by titrating the remaining unreacted acid. From this result, the original nitrogen content of the material can be calculated. This method was developed by Danish chemist Johan Kjeldahl (1849–1900).
Separation techniques and infrared spectroscopy-based methods are also used for casein measurement, but these solutions are quite complex (*Dimenna et al*., 1981; *Bonfatti et al*., 2008). Therefore, the data-driven approach I developed is significantly simpler for estimating casein content and can be highly beneficial in the dairy industry.

Further development of the model (into a time-based prediction tool) could improve breeding efficiency. The effectiveness of this solution demonstrates that machine learning models can help predict chemical and biological values in milk. The advantage of this method is that it builds the model by processing existing databases, eliminating the need for large investments. Since data collection does not require new investments or different sampling techniques, my developed system can be easily integrated into any production method.

10. MELLÉKLETEK

M1. Irodalomjegyzék

1. ALAM, M. et al. (2015): Estimation of Genetic Parameters for Somatic Cell Scores of Holsteins Using Multi-trait Lactation Models in Korea. In: *Asian-Australasian Journal of Animal Sciences,* 28.3, 303–310. p. DOI: 10.5 713/ajas.13.0627.

2. ALFÖLDI L.; TARR Z.; TŐZSÉR J. (2020): Digitális mikroklíma mérés a tejtermelő farmon. In: *Animal Welfare, Etológia és Tartástechnológia (AWETH),* 16 (2) 94–109. p. DOI: 10.17205/szie.aweth.2020.2.094.

3. ALI, I. et al. (2017): Modeling Managed Grassland Biomass Estimation by Using Multitemporal Remote Sensing Data—A Machine Learning Approach. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10 (7) 3254–3264 p. DOI: 10.1109/JSTARS.2016.2561618.

4. ALONSO, J.; VILLA, A.; BAHAMONDE, A. (2015): Improved estimation of bovine weight trajectories using Support Vector Machine Classification. In: *Computers and Electronics in Agriculture,* 110, 36–41. p. DOI: 10.1016/j.compag.2014.10.001.

5. ALVAREZ, J. R. et al. (2018): Body condition estimation on cows from depth images using Convolutional Neural Networks. In: *Computers and Electronics in Agriculture,* 155, 12–22. p., ISSN 0168-1699, DOI: 10.1016/j.compag.2018.09.039.

6. ALVAREZ, J.; ARROQUIA, M.; MANGUDOA, P. (2018): Body condition estimation on cows from depth images using Convolutional Neural Networks. In: *Computers and Electronics in Agriculture,* 155, 12–22. p. DOI: 10.1016/j.compag.2018.09.039.

7. ALVAREZ, J.; ARROQUIA, M.; MANGUDOA, P. (2019): Estimating Body Condition Score in Dairy Cows From Depth Images Using Convolutional Neural Networks, Transfer Learning and Model Ensembling Techniques. In: *Agronomy* 9 (2), 90. p. DOI: 10.3390/agronomy9020090.

8. AMIN, A. (2018): Estimates of Heritability for Somatic Cell Count, Test-Day Milk Yield and Some Udder-Teat Characteristics in Saudi Dairy Goats using Random Regression Animal Model. In: *Advances in Animal and Veterinary Sciences,* 6. 128–134. p. DOI: 10.17582/journal.aavs/2018/6.3.128.134.

9. ASHWINI, J. et al. (2019): Prediction of Body Weight based on Body Measurements in Crossbred Cattle. In: *International Journal of Current Microbiology and Applied Sciences*, 8 (3) 1597–1611. p.

10. BARBEDO, J. G. A. et al. (2019): A Study on the Detection of Cattle in UAV Images Using Deep Learning. In: *Sensors* 19 (24), 5436. p. DOI: 10.3390/s19245436.

11. BARBEDO, J. G. A. et al. (2020): Counting Cattle in UAV Images—Dealing with Clustered Animals and Animal/Background Contrast Changes. In: *Sensors,* 20 (7). 2126. p. DOI: 10.3390/s20072126.

12. BARRIUSO, A. L. et al. (2018): Combination of Multi-Agent Systems and Wireless Sensor Networks for the Monitoring of Cattle. In: *Sensors,* 18 (1), 108. p. DOI: 10.3390/s18010108.

13. BHAT, M. Y.; Dar, T. A.; Singh, L. R. (2016): Casein Proteins: Structural and Functional Aspects. In: *InTech.* DOI: 10.5772/64187.

14. BISUTTI, V. et al. (2022): Impact of somatic cell count combined with differential somatic cell count on milk protein fractions in Holstein cattle. In: *Journal of Dairy Science,* 105 (8) 6447–6459. p. DOI: 10.3168/jds.2022-22071.

15. BONFATTI, V. et al. (2008): Validation of a new reversed-phase high-performance liquid chromatography method for separation and quantification of bovine milk protein genetic variants. In: *Journal of Chromatography A,* 1195 (1–2), 101–106. p. DOI: 10.1016/j.chroma.2008.04.075.

16. BORECKI, M. et al. (2009): A method of testing the quality of milk using optical capillaries. In: *Photonics Letters of Poland,* 1, 37–39. p.

17. BOUDRON, R. M.; BRINKS, J. S. (1986): C54 Beef Program Report, Colorado State University, April, 52–57. p.

18. BUCCOLA, S.; IIZUKA, Y. (1997): Hedonic Cost Models and the Pricing of Milk Components. In: *American Journal of Agricultural Economics,* 79 (2) 452–462. p. DOI: 10.2307/1244143.

19. BURVENICH, C. P. G. et al. (2000): Physiological and Genetic Factors That Influence the Cows Resistance to Mastitis, Especially during Early Lactation. In: *Proceedings of the 5th IDF Mastitis Congress, Symposium on Immunology of Ruminant Mammary Gland,* 9–20. p.

20. CHENG, J. B. et al. (2007): Factors Affecting the Lactoferrin Concentration in Bovine Milk. In: *Journal of Dairy Science,* 91 (3) 970–976. p. DOI: 10.3168/jds.2007-0689.

21. CHICCO, D.; WARRENS, M. J.; JURMAN, G. (2021): The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. In: *PeerJ. Computer science,* 7 (e623), Published 2021 Jul 5. DOI: 10.7717/peerj-cs.623. PMID: 34307865.

22. CHLEBOWSKI, J. et al. (2020): Association between body condition and production parameters of dairy cows in the experiment with use of BCS camera. In: *Agronomy Research,* 18 (S2), 1203–1212. p. DOI: 10.15159/ar.20.028.

23. CHUNG, C. L. et al. (2016): Detecting Bakanae disease in rice seedlings by machine vision. In: *Computers and Electronics in Agriculture,* 121, 404–411. p. DOI: 10.1016/j.compag.2016.01.008.

24. CRANINX, M. et al. (2008): Artificial neural network models of the rumen fermentation pattern in dairy cattle. In: *Computers and Electronics in Agriculture,* 60, 226–238. p. DOI: 10.1016/j.compag.2007.08.005.

25. DABDOUB, S.; SHOOK, G. (1984): Phenotypic relations among milk yield, somatic cell count and clinical mastitis. In: *Journal of Dairy Science,* 67 (1), 163–164. p.

26. DÉGEN L. et al. (2018): Partnertájékoztató Hírlevél, az Állattenyésztési Teljesítményvizsgáló Kft. kiadványa, ISSN HU-2063-3491, 2018. XVIII. évfolyam 1. szám.

27. DIMENNA, G.; SEGALL, H. (1981): High-performance gel-permeation chromatography of bovine skim milk proteins. In: *Journal of Liquid Chromatography,* 4 (4), 639–649. p. DOI: 10.1080/01483918108059961.

28. DOHOO, I.; LESLIE, K. E. (1991): Evaluation of changes in somatic cell counts as indicators of new intramammary infections. In: *Preventive*

Veterinary Medicine, 10 (3) 225–237. p. DOI: 10.1016/0167-5877(91)90006-N.

29. DUTTA, R. et al. (2015): Dynamic cattle behavioural classification using supervised ensemble classifiers. In: *Computers and Electronics in Agriculture,* 111, 18–28. p. DOI: 10.1016/j.compag.2014.12.002.

30. EBRAHIMI, M. A. et al. (2017): Vision-based pest detection based on SVM classification method. In: *Computers and Electronics in Agriculture,* 137, 52–58. p. DOI: 10.1016/j.compag.2017.03.016.

31. FERENTINOS, K. P. (2018): Deep learning models for plant disease detection and diagnosis. In: *Computers and Electronics in Agriculture,* 145, 311–318. p. DOI: 10.1016/j.compag.2018.01.009.

32. FERNÁNDEZ-CARRIÓN, E. et al. (2017): Motion-based video monitoring for early detection of livestock diseases: The case of African swine fever. In: *PloS one,* 12 (9), DOI: 10.1371/journal.pone.0183793.

33. FRANCIS, J.; SIBANDA, S.; KRISTENSEN, T. (2002): Estimating Body Weight of Cattle Using Linear Body Measurements. In: *Zimbabwe Veterinary Journal,* 33 (1) 15–21. p. DOI: 10.4314/zvj.v33i1.5297.

34. FRASER, Evan D. G. (2020): The challenge of feeding a diverse and growing population. In: *Physiology & Behavior,* 221, 112908, ISSN 0031-9384, DOI: 10.1016/j.physbeh.2020.112908.

35. GARCÍA, R. et al. (2020): A systematic literature review on the use of machine learning in precision livestock farming. In: *Computers and Electronics in Agriculture,* 179, 105826. DOI: 10.1016/j.compag.2020.105826.

36. GEURTS, P.; ERNST, D.; WEHENKEL, L. (2006): Extremely randomized trees. In: *Machine Learning,* 63, 3–42. p. DOI: 10.1007/s10994-006-6226-1.

37. GRINBLAT, G. L. et al. (2016): Deep learning for plant identification using vein morphological patterns. In: *Computers and Electronics in Agriculture,* 127, 418–424. p. DOI: 10.1016/j.compag.2016.07.003.

38. HALASA, T.; KIRKEBY, C. (2020): Differential Somatic Cell Count: Value for Udder Health Management. In: *Frontiers in Veterinary Science,* 7, DOI: 10.3389/fvets.2020.609055.

39. HANSEN, M. F. et al. (2018): Towards on-farm pig face recognition using convolutional neural networks. In: *Computers in Industry,* 98, 145–152. p. DOI: 10.1016/j.compind.2018.02.016.

40. HOLLÓSI D. (szerk: MILICS G.) (2017): Adatalapú döntések a 2020 utáni finanszírozásban. Precíziós Gazdálkodás, Adat, Információ, Haszon. Budapest, Agroinform és NAK, 26. p.

41. HU, H. et al. (2017): Differentiation of deciduous-calyx and persistent-calyx pears using hyperspectral reflectance imaging and multivariate analysis. In: *Computers and Electronics in Agriculture,* 137, 150–156. p. DOI: 10.1016/j.compag.2017.04.002.

42. JOURNAUX, L. (1994): Description et évaluation génétique de la morphologie des bovins allaitants en France. Performance recording of animals: State of the Art. In: *EAAP Publication,* 32–48. p., https://hal.inrae.fr/hal-02712092v1.

43. KAWASAKI, M. et al. (2008): Near-infrared spectroscopic sensing system for on-line milk quality assessment in a milking robot. In: *Computers and Electronics in Agriculture,* 63 (1), 22–27. p. DOI: 10.1016/j.compag.2008.01.006.

44. KHOSIA, R. (szerk: MILICS G.) (2017): A precíziós gazdálkodás és az adatok forradalma a nagy és kis méretű mezőgazdasági rendszerekben. Precíziós Gazdálkodás, Adat, Információ, Haszon. Budapest, Agroinform és NAK, 18. p.

45. KNIGHTS, S. A. et al. (1984): Estimation of heritabilities and of genetic and phenotypic correlation among growth and reproductive traits in yearling Angus bulls. In: *Journal of animal science*, 58 (4) 887–893. p. DOI: 10.2527/jas1984.584887x.

46. KOCSIS R. et al. (2022): Annual and seasonal trends in cow's milk quality determined by FT-MIR spectroscopy in Hungary between 2011 and 2020. In: *Acta Veterinaria Hungarica,* 70 (3), 207–214. p. DOI: 10.1556/004.2022.00019.

47. KODORS, S. et al. (2021): Apple scab detection using CNN and Transfer Learning. In: *Agronomy Research,* 19 (2), 507–519. p. DOI: 10.15159/AR.21.045.

48. KOLLÁR Cs.; NAGY B. (2021): A mesterséges intelligencia felhasználási lehetőségei az objektumfelismerésben. In: *Biztonságtudományi Szemle,* 3 (1) 123–140. p.

49. KORCHMA, Cs. (1986): Eltérő technológiával hizlalt, különböző genotípusú növendékbikák vágási és küllemi értékmérőinek összefüggés-vizsgálata a húshasznú tenyészbikák szelekciós rendszerének korszerűsítése érdekében. Doktori értekezés, Gödöllő, Agrártudományi Egyetem, 1–225. p.

50. KUMAR, S.; SINGH, S. K. (2017): Automatic identification of cattle using Fzzmuzzle point pattern: a hybrid feature extraction and classification paradigm. In: *Multimedia Tools and Applications,* 76, 26551–26580. p. DOI: 10.1007/s11042-016-4181-9.

51. KUNES, R. et al. (2021): In-Line Technologies for the Analysis of Important Milk Parameters during the Milking Process: A Review. In: *Agriculture,* 11 (3) 239. p. DOI: 10.3390/agriculture11030239.

52. KUNG, H.-Y.et al. (1998): Accuracy Analysis Mechanism for Agriculture Data Using the Ensemble Neural Network Method. In: *Sustainability,* 8 (8) 735. DOI: 10.3390/su8080735.

53. KÜHL, H. S.; BURGHARDT, T. (2013): Animal biometrics: quantifying and detecting phenotypic appearance. In: *Trends in ecology & evolution*, 28 (7) 432–441. p. DOI: doi.org/10.1016/j.tree.2013.02.013.

54. LINKO, S. (1998): Expert systems—what can they do for the food industry? In: *Trends in Food Science and Technology,* 9 (1) 3–12. p. DOI: 10.1016/S0924-2244(97)00002-2.

55. LUKUYU M. N. et al. (2016): Use of body linear measurements to estimate liveweight of crossbred dairy cattle in smallholder farms in Kenya. In: *Springer Plus,* 5, 1–14. p. DOI: 10.1186/s40064-016-1698-3.

56. MAIONE, C. et al. (2016): Classification of geographic origin of rice by data mining and inductively coupled plasma mass spectrometry. In: *Computers and Electronics in Agriculture,* 121, 101–107. p. DOI: 10.1016/j.compag.2015.11.009.

57. MATTHEWS, S. G. et al. (2017): Automated tracking to measure behavioural changes in pigs for health and welfare monitoring. In: *Scientific Reports,* 7, 17582. DOI: 10.1038/s41598-017-17451-6.

58. MILLER, G. et al. (2019): Using 3D Imaging and Machine Learning to Predict Liveweight and Carcass Characteristics of Live Finishing Beef Cattle Front. In: *Frontiers in Sustainable Food Systems,* 3, DOI: 10.3389/fsufs.2019.00030.

59. MORALES, I. R. et al. (2016): Early warning in egg production curves from commercial hens: A SVM approach. In: *Computers and Electronics in Agriculture,* 121, 169–179. p. DOI: 10.1016/j.compag.2015.12.009.

60. MOROTA, G. et al. (2018): BIG DATA ANALYTICS AND PRECISION ANIMAL AGRICULTURE SYMPOSIUM: Machine learning and data mining advance predictive big data analysis in precision animal agriculture. In: *Journal of Animal Science*, 96 (4), 1540–1550. p. DOI: 10.1093/jas/sky014.

61. MOSHOU, D. et al. (2014): Water stress detection based on optical multisensor fusion with a least squares support vector machine classifier. In: *Biosystems Engineering,* 117, 15–22. p. DOI: 10.1016/j.biosystemseng.2013.07.008.

62. MUHAMMAD, W. A.; REYNOLDS, J.; REZGUI, Y. (20187): Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. In: *Journal of Cleaner Production,* 203, 810–821. p. DOI: 10.1016/j.jclepro.2018.08.207.

63. NANDAN, D.; KUMAR, P. (2022): An Overview on Properties and Constituents of Cow's Milk. In: *International Journal of Innovative Research in Engineering & Management* (IJIREM), ISSN: 2350-0557, 9 (1), Article ID IRP229177, 380–383. p. DOI: 10.55524/ijirem.2022.9.1.79.

64. NEELY, J. D. et al. (1982): Genetic parameters for testis size and sperm number in Hereford bulls. In: *Journal of animal science,* 55 (5) 1033–1040. p. DOI: 10.2527/jas1982.5551033x.

65. NEETHIRAJAN, S. (2020): The role of sensors, Big Data and machine learning in modern animal farming. In: *Sensing and Bio-Sensing Research,* 29, 100367. DOI: 10.1016/j.sbsr.2020.100367.

66. NEETHIRAJAN, S.; KEMP, B. (2021): Digital Livestock Farming. In: *Sensing and Bio-Sensing Research,* 32. DOI: 10.1016/j.sbsr.2021.100408.

67. NICKERSON, S. C. (1995): Milk production: Factors affecting milk composition. In: *Harding, F. (eds): Milk Quality. Springer, Boston, MA.* DOI: 10.1007/978-1-4615-2195-2_2.

68. OZKAYA, S.; BOZKURT, Y. (2009): The accuracy of prediction of body weight from body measurements in beef cattle. In: *Archiv für Tierzucht,* 52 (4), 371–377. p. DOI: 10.5194/aab-52-371-2009.

69. PANTAZI, X. E. et al. (2016): Wheat yield prediction using machine learning and advanced sensing techniques. In: *Computers and Electronics in Agriculture,* 121, 57–65. p. DOI: 10.1016/j.compag.2015.11.018.

70. PANTAZI, X. E. et al. (2017): Detection of biotic and abiotic stresses in crops by using hierarchical self organizing classifiers. In: Precision Agriculture, 18, 1–11. p. DOI: 10.1007/s11119-017-9507-8.

71. PANTAZI, X. E. et al. (2017): Detection of Silybum marianum infection with Microbotryum silybum using VNIR field spectroscopy. In: *Computers and Electronics in Agriculture,* 137, 130–137. p. DOI: 10.1016/j.compag.2017.03.017.

72. PANTAZI, X. E. et al. (2017): Evaluation of hierarchical self-organising maps for weed mapping using UAS multispectral imagery. In: *Computers and Electronics in Agriculture,* 139, 224–230. p. DOI: 10.1016/j.compag.2017.05.026.

73. PANTAZI, X. E.; MOSHOU, D.; BRAVO, C. (2016): Active learning system for weed species recognition based on hyperspectral sensing. In: *Biosystems Engineering,* 146, 193–202. p. DOI: 10.1016/j.biosystemseng.2016.01.014.

74. PEGORINI, V. et al. (2015): In vivo pattern classification of ingestive behavior in ruminants using FBG sensors and machine learning. In: *Sensors*, 15, 28456–28471. p. DOI: 10.3390/s151128456.

75. QIAO, Y. et al. (2019): Individual Cattle Identification Using a Deep Learning Based Framework. In: *IFAC-PapersOnLine,* 52 (30) 318–323. p. DOI: 10.1016/j.ifacol.2019.12.558.

76. RAMOS, P. J. et al. (2017): Automatic fruit count on coffee branches using computer vision. In: *Computers and Electronics in Agriculture,* 137, 9–22. p. DOI: 10.1016/j.compag.2017.03.010.

77. Research for AGRI Committee (2019): Impacts of the digital technology on the food chain and the CAP. European Parliament, Policy Department for Structural and Cohesion Policies, Brussels.

78. RODRÍGUEZ ALVAREZ, J. et al. (2019): Estimating Body Condition Score in Dairy Cows From Depth Images Using Convolutional Neural Networks, Transfer Learning and Model Ensembling Techniques. In: *Agronomy,* 9 (2) 90. p. DOI: 10.3390/agronomy9020090.

79. RUSSELL, S.; NORVIG, P. (2005): Artificial Intelligence: A Modern Approach.
2nd Edition, Pearson Education Limited, London.

80. RUSSELL, S.; NORVIG, P. (2021): Artificial Intelligence: A Modern Approach. Global Edition, Pearson Education Limited, London.

81. RUTTER, M. (2011): What sensors work for livestock now and where might we go? Harper Adams University, The National Centre for Precision Farming.

82. SADEGHI, M. et al. (2015): An Intelligent Procedure for the Detection and Classification of Chickens Infected by Clostridium Perfringens Based on their Vocalization. In: *Revista Brasileira de Ciência Avícola,* 17, 537–544. p. DOI: 10.1590/1516-635x1704537-544.

83. SEABOLD, S.; PERKTOLD, J. (2010): Statsmodels: Econometric and Statistical Modeling with Python. Proceedings of the 9th Python in Science Conference.

84. SENGUPTA, S.; LEE, W. S. (2014): Identification and determination of the number of immature green citrus fruit in a canopy under different ambient light conditions. In: *Biosystems Engineering,* 117, 51–61. p. DOI: 10.1016/j.biosystemseng.2013.07.007.

85. SENTHILNATH, J. et al. (2016): Detection of tomatoes using spectral-spatial methods in remotely sensed RGB images captured by UAV. In: *Biosystems Engineering,* 146, 16–32. p. DOI: 10.1016/j.biosystemseng.2015.12.003.

86. SHAHINFAR, S. et al. (2012): Prediction of Breeding Values for Dairy Cattle Using Artificial Neural Networks and Neuro-Fuzzy Systems. In: *Computational and Mathematical Methods in Medicine*, 2012 (1) 127130. DOI: 10.1155/2012/127130.

87. SHARMA, A. et al. (2020): Machine Learning Applications for Precision Agriculture: A Comprehensive Review. In: *IEEE Access*, 9, 4843–4873. p. DOI: 10.1109/ACCESS.2020.3048415.

88. SHARMA, N.; SINGH, N. K.; BHADWAL, M. S. (2011): Relationship of Somatic Cell Count and Mastitis: An Overview. In: *Asian-Australasian Journal of Animal Sciences,* 24, 429–438. p. DOI: 10.5713/ajas.2011.10233.

89. SØRENSEN, L. K.; LUND, M.; JUUL, B. (2003): Accuracy of Fourier transform infrared spectrometry in determination of casein in dairy cows' milk. In: *Journal of Dairy Research,* 70 (4), 445–452. p. DOI: 10.1017/S0022029903006435.

90. STAFFORD, J. (szerk: MILICS G.) (2017): Precíziós gazdálkodás, Big Data és UAV-k – képzelet vagy valóság? Precíziós Gazdálkodás, Adat, Információ, Haszon. Budapest, Agroinform és NAK, 19. o.

91. SU, Y.; XU, H.; YAN, L. (2017): Support Vector Machine-Based Open Crop Model (SBOCM): Case of Rice Production in China. In: *Saudi Journal of Biological Sciences,* 24, 537–547. p. DOI: 10.1016/j.sjbs.2017.01.024.

92. SWAISGOOD, H. E. (1982): Chemistry of milk protein. In: Fox P. F. (ed.): Developments in Dairy Chemistry. Elsevier Applied Science Publishers, London, UK, 1982, 1–59. p.

93. SZABÓ I. (2019): Az agrárinformatika helyzete, fejlődési irányai hazánkban és nemzetközi kitekintésben. In: *Állattenyésztés és takarmányozás,* 68 (3) 185–194. p.

94. TEBUG, S. et al. (2018): Using body measurements to estimate live weight of dairy cattle in low-input systems in Senegal. In: *Journal of Applied Animal Research*, 46 (1) 87–93. p. DOI: 10.1080/09712119.2016.1262265.

95. TELES Jr., C. G. S. et al. (2019): A software to estimate heat stress impact on dairy cattle productive performance. In: *Agronomy Research,* 17 (3), 872–878. p. DOI: 10.15159/AR.19.110

96. TÍMÁR I. (2004): Versenyképesség a magyar tejágazatban. Doktori értekezés, Budapesti Corvinus Egyetem.

97. TŐZSÉR J.; SZŰCS M. (2020): Regressziós elemzések a szelekciós célok pontosítására a központi sajátteljesítmény-vizsgálatban a Limousin fajtában. In: *Animal Welfare, Etológia és Tartástechnológia (AWETH)*, 16 (2) 189–199. p. DOI: 10.17205/SZIE.AWETH.2020.2.189.

98. VAN DER VEN, C. et al. (2001): Reversed phase and size exclusion chromatography of milk protein hydrolysates: relation between elution from reversed phase column and apparent molecular weight distribution. In: *International Dairy Journal,* 11 (1–2) 83–92. p. DOI: 10.1016/S0958-6946(01)00032-2.

99. VANDERWAAL, K. et al. (2017): Translating Big Data into Smart Data for Veterinary Epidemiology. In: *Frontiers in Veterinary Science,* 4. DOI: 10.3389/fvets.2017.00110.

100. WEBER, F. et al. (2020): Recognition of Pantaneira cattle breed using computer vision and convolutional neural networks. In: *Computers and Electronics in Agriculture,* 175. 10. DOI: 1016/j.compag.2020.105548.

101. WEBER, V. A. et al. (2020): Prediction of Girolando cattle weight by means of body measurements extracted from images. In: *Revista Brasileira de Zootecnia,* 49. DOI: 10.37496/rbz4920190110.

102. XU, B. et al. (2020): Livestock classification and counting in quadcopter aerial images using Mask R-CNN. In: *International Journal of Remote Sensing,* 41 (21), 8121–8142. p. DOI: 10.1080/01431161.2020.1734245.

103. ZHANG, M.; LI, C.; YANG, F. (2017): Classification of foreign matter embedded inside cotton lint using short wave infrared (SWIR) hyperspectral transmittance imaging. In: *Computers and Electronics in Agriculture,* 139, 75–90. p. DOI: 10.1016/j.compag.2017.05.005.

M2. Az értekezés témaköréhez kapcsolódó saját publikációk

1. Revoly A.; Tarr B.; Szabó I. (2023): A mesterséges intelligencia szerepe az élelmiszerbiztonságban. In: *Biztonságtudományi Szemle,* 5 (3) 141–150. p.

2. Szabó I.; Tarr B.; Revoly A. (2023): Adattechnológiák implementációja mezőgazdasági műszaki folyamatokban. Előadás, MTA Tudományos ülés, Budapest, 2023. 11. 06.

3. Tarr B. et al. (2022): Precíziós eljárások és a mesterséges intelligencia technológia alkalmazása a szarvasmarha-tenyésztésben különös tekintettel a húshasznú szarvasmarhák azonosítására. In: *Animal welfare Etológia és Tartástechnológia (AWETH),* 18 (1) 51–63. p.

4. Tarr B. et al. (2023): Predicting somatic cell count in milk samples using machine learning. In: *12th International Conference on Applied Informatics (ICAI 2023): Abstracts.* 1–2. p.

5. Tarr B.; Szabó I.; Tőzsér J. (2023): Estimation of important milk constituents from milk samples using artificial intelligence. In: *5th International Conference on Biosystems and Food Engineering: Proceedings.*

6. Tarr B.; Szabó I.; Tőzsér J. (2023): Machine learning in cattle breeding. In: *Mechanical Engineering Letters,* 24, 71–84. p.

7. Tarr B.; Szabó I.; Tőzsér J. (2023): Mesterséges intelligencia alkalmazása a szarvasmarha-tenyésztés egyes területein: egyedazonosítás, állategészségügy és állatjóllét: Irodalmi összefoglaló. In: *Magyar Állatorvosok Lapja,* 145 (11) 651–660. p.

8. Tarr B.; Szabó I.; Tőzsér J. (2024): Predicting somatic cell count in milk samples using machine learning. In: *Annales Mathematicae et Informaticae,* 60, 159–168. p.

9. Tarr B.; Szabó I.; Tőzsér J. (2025): Body Conformation Scoring of Cattle, using Machine Learning. In: *Acta Polytechnica Hungarica,* 22 (3) 27–38. p.

10. Tarr B.; Tőzsér J.; Szabó I. (2021): Precíziós gazdálkodás módszerének és a mesterséges intelligencia (MI) eljárásának alkalmazási lehetőségei a szarvasmarha-tenyésztésben. In: *Mezőgazdasági Technika,* 63 (7) 2–5. p.

11. Tőzsér J. et al. (2023): Evaluation of body measurements of Limousin heifers by backward regression analysis in western Hungary. In: *Animal welfare Etológia és Tartástechnológia (AWETH),* 19 (1) 102–117. p.