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 tézisei

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

1. A MUNKA ELŐZMÉNYEI, CÉLKITŰZÉSEK 2

1.1. A SZAKIRODALOM KRITIKAI ELEMZÉSE 4

2. ANYAG ÉS MÓDSZER 5

2.1. MÓDSZEREK 6

2.1.1. Lineáris regresszió 7

2.1.2. Poisson-regresszió 8

2.1.3. A Regressziós modellek értékelési mutatói 8

2.1.4. Random Forest 9

2.1.5. Decision Tree Regressor 9

2.1.6. Extra Trees Classifier 9

2.1.7. Support Vector Machine 10

2.1.8. Döntési fák kiértékelési módszerei 11

2.1.9. Hagyományos matematikai módszerek 11

2.1.10. Kiegyensúlyozatlan adathalmazok 12

3. EREDMÉNYEK ÉS AZOK MEGBESZÉLÉSE 13

3.1. KÜLLEMI BÍRÁLATI PONTSZÁMOK BECSLÉSE 13

3.2. SZOMATIKUS SEJTSZÁM BECSLÉSE 16

3.3. A KAZEIN MENNYISÉGÉNEK BECSLÉSE 20

4. KÖVETKEZTETÉSEK ÉS JAVASLATOK 24

5. ÚJ TUDOMÁNYOS EREDMÉNYEK 25

5.1. 1. TÉZIS 25

5.2. 2. TÉZIS 27

5.3. 3. TÉZIS 30

5.4. 4. TÉZIS 31

6. AZ ÉRTEKEZÉS TÉMAKÖRÉHEZ TARTOZÓ SAJÁT PUBLIKÁCIÓK 33

1. A MUNKA ELŐZMÉNYEI, CÉLKITŰZÉSEK

A mesterséges intelligencia használata az állattenyésztés területén is meghatározó feltétele a további mennyiségi és minőségi növekedésnek. Az utóbbi évtizedekben számos kutatás foglalkozott a mesterséges intelligencia alkalmazási lehetőségeivel az állattenyésztés területén is. 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, és támogatják a hatékonyabb tenyésztést. 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 kicsi beruházási igénnyel lenne használható a tenyésztők számára a fent említett témákban.

1.1. A szakirodalom kritikai elemzése

Az általam áttekintett 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.

2. 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. A modellem tanításához a *Limousin és Blonde d'Aquitaine Tenyésztők Egyesületétől* (Budapest) kapott 325 állat pontozási eredményét tartalmazó adatbázist használtam.

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. 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.

2.1. 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.
A következőkben áttekintem a kutatásom során használt algoritmusokat.

2.1.1. 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:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ij} + \varepsilon_i \,, \qquad (2.1)$$

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, $\varepsilon_i \sim N(0, \sigma^2)$).

2.1.2. 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).

2.1.3. A Regressziós modellek értékelési mutatói

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.

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

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.

2.1.4. 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.

2.1.5. 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.

2.1.6. 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.

2.1.7. 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.

2.1.8. Döntési fák kiértékelési módszerei

Pontosság (*Accuracy*)
Azt méri, hogy a modell hány esetben adott helyes predikciót az összes minta közül.

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.

2.1.9. 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.

2.1.10. 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.

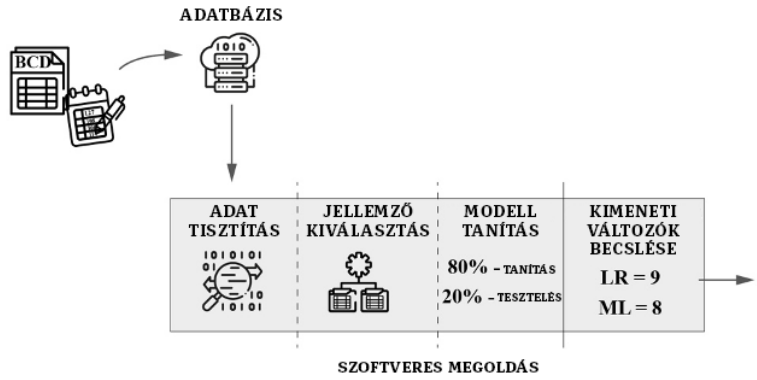
A kiegyensúlyozatlan adathalmazok kiegyensúlyozására (adatszintű megközelítésben) alapvetően kétféle módszert alkalmazhatunk. Az egyik módszer a túlreprezentált adatok elhagyása és további alulreprezentált adatok beszerzése. A másik módszer az alulreprezentált adatok többszörözése (SMOTE – *Synthetic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

3. EREDMÉNYEK ÉS AZOK MEGBESZÉLÉSE

3.1. 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 *3.1. ábrán* látható.



3.1. ábra: A becslő szoftver működésének elvi vázlata

Az adatbázisban tárolt korábbi értékelési adatokat *(3.1. 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.

3.1. 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 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 *3.2. táblázatban* látható.

3.2. 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 az 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 kedvezőek 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 jól előre jelezhetők a többi három jellemző eredményei ($R^2 = 0,86$).

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.

3.2. Szomatikus sejtszám becslése

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.

Az adatbázisban tárolt adatokat megvizsgálva a szomatikus sejtszám változó statisztikai elemzése az *3.3. táblázatban* látható.

3.3. 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.

Az SCC értékeit három csoportba osztottam (nem fertőzött = 0, lehetséges fertőzés = 1, fertőzött = 2), és ezt a három értéket fogom használni a minták osztályozásához.
A kimeneti változó egyes kategóriának az előfordulását az *3.4. táblázat* tartalmazza.

3.4. 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 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 legjobb eredményt adó modell a következő bemeneti változókat használta: LNPC, karbamid, cukor, laktoferrin, zsír és fehérje. Az egyes modellek átlagos pontosságát az *3.5. táblázat* tartalmazza.

3.5. 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.

3.3. A kazein mennyiségének becslése

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 ezért egy regresszióalapú modell használata mellett döntöttem.

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 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).

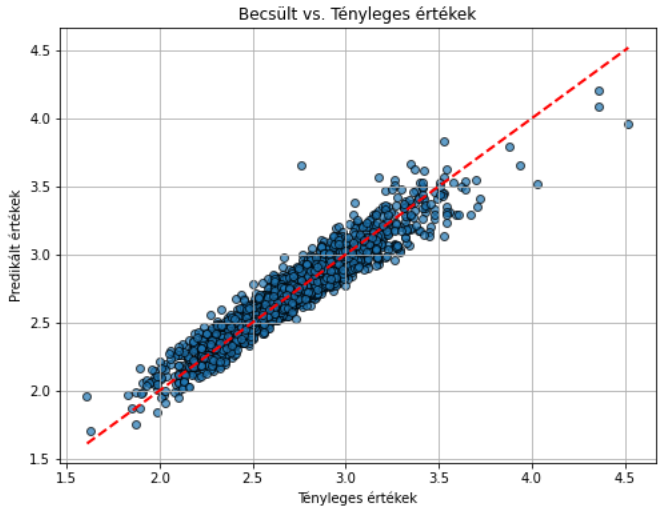
Az eredmények ellenőrzésére az R^2 determinációs együtthatót és az átlagos négyzetes gyökeltérést (RMSEP) használtam.

10 sorozatot (tanítás–ellenőrzés) futtattam minden egyes bemeneti csoporttal úgy, hogy az adatbázist minden alkalommal véletlenszerűen osztottam fel tanító és ellenőrző részekre. Az *3.6. táblázat* a futások legjobb és legrosszabb eredményeit tartalmazza.

3.6. 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 *(3.2. ábra)*. A piros vonal jelzi az egyes kimeneti értékekből képzett egyenest, a kék pöttyök pedig a becsült értékeket.



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

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.

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.

4. 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.

5. 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.

5.1. 1. Tézis

1. Küllemi bírálati pontok előrejelzése

Kutatásom során létrehoztam és validáltam egy új 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 *5.1. táblázatban* láthatjuk.

5.1. 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

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.

5.2. 2. Tézis

2. A szomatikus sejtszám előrejelzése

Kutatásom során bizonyítottam, hogy lehetséges olyan, gépi tanuláson alapuló új 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 új modellt hoztam létre 25 000 tejminta adatai alapján, ami más, mérhető tejparamétereket használva előre jelzi a szomatikus sejtszám mértékét.

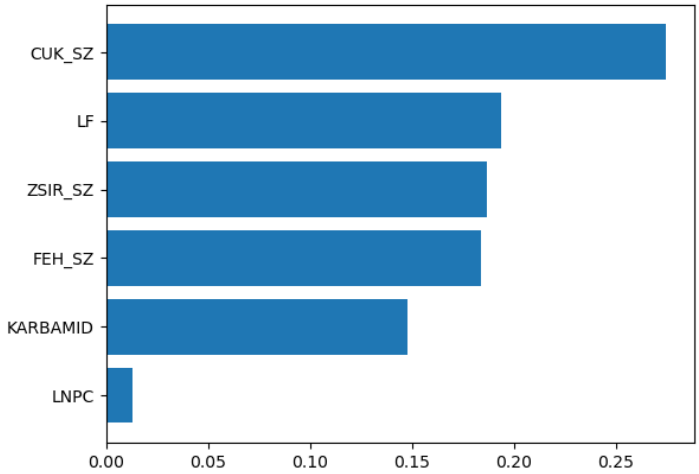
A kutatás során bebizonyítottam, hogy más tejparaméterek segítségével javíthatom a masztitisz korai felismerésének lehetőségét, miközben a gépi tanulás alkalmazása csökkenti a laboratóriumi módszerek költségeit és időigényét.

Az általam vizsgált algoritmusok közül az *Extra Trees Classifier* algoritmussal tanított modell hozta a legjobb eredményeket, amelynek eredményeit az 5.*2. táblázatban* láthatjuk.

5.2. 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 *5.1. ábrán* láthatjuk.



5.1. 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]

5.3. 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.

5.3. 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.

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]

5.4. 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.

6. AZ ÉRTEKEZÉS TÉMAKÖRÉHEZ TARTOZÓ 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.