



Factors affecting the longevity of Holstein dairy cattle: A data mining approach

Ali Rezazadehvishkaei¹ | Alireza Hasanibafarani^{2✉} | Kian Pahlevanafshar³ | Mehran Aboozari⁴

1. Department of Animal Science, Faculty of Agriculture, Abhar Branch, Islamic Azad University, Zanjan, Iran. E-mail: ali.rezazadeh@iau.ac.ir.
2. Corresponding Author, Department of Animal Science, Agricultural Research, Education and Extension Organization (AREEO), Agricultural institute of education and extension (IATE), Tehran, Iran. E-mail: arhb@abc.org.ir
3. Department of Animal Science, Faculty of Agriculture, Varamin Branch, Islamic Azad University Tehran, Iran. E-mail: pahlevanafshar.k@abhariau.ac.ir
4. Department of Animal Science, Faculty of Agriculture, Abhar Branch, Islamic Azad University. Zanjan, Iran. E-mail: m.aboozari1357@iau.ac.ir

Article Info

Article type:
Research Article

Article history:

Received 26 October 2024
Received in revised form
1 June 2025
Accepted 3 June 2025
Published online 11 October 2025

Keywords:

Longevity
Dairy cattle lifespan
Reproductive traits
Machine learning
Classification algorithms

ABSTRACT

Objective: This study seeks to identify and systematically analyze key factors affecting the longevity of Holstein dairy cows in the herds, using advanced data mining algorithms. Understanding and predicting longevity is vital because it directly impacts productivity and profitability of dairy farm. Longer-lasting cows tend to produce more calves and produce more milk, thereby increasing the overall economic efficiency of dairy operations. Furthermore, extended longevity is associated with reduced replacement costs.

Methods: In recent years, the integration of machine learning techniques into agricultural and livestock management has gained significant momentum. This study uses detailed phenotypic data collected from 37,009 female animals belonging to 664 sires in 82 separate herds, representing a comprehensive dataset spanning a decade. The data includes eight milk production records, along with other relevant variables such as animal age, sire number, shelf life in months, somatic cell count, days in milk, milk production (kg), protein and fat content, calving cycle length, milking frequency, geographic location (province), birth date, calving date, calving interval, herd code, and age at first calving. The data preparation phase included processing and organizing the dataset using Excel 2016, which ensured data quality and consistency. Subsequent data analyses were conducted using R software (version 4.3.3), and using the relevant specialized packages for machine learning and statistical modeling.

Results: The results showed that, the Support Vector Machine has the best accuracy (0.987). Random Forest was the second most efficient algorithm. The accuracy of the Gradient Boosting Machine was slightly lower than that of the Random Forest, but it still showed good performance. The Decision Tree provided the least accuracy among these algorithms. The Decision Tree and Support Vector Machine achieved this performance with fewer input variables compared to Gradient Boosting Machine and Random Forest.

Conclusion: The results showed that none of the algorithms used for survival classification, despite acceptable accuracy, are error-free, but on the other hand, it was shown that the decision tree is simpler and less expensive. The most important features of these methods are the lack of statistical assumptions and requirements that linear regression and interpolation methods require, the lack of normality assumptions, robustness against missing values and the ability to detect complex nonlinear relationships between variables and prediction targets, which makes them suitable for various applications in the livestock industry. Accurate data recording protocols as well as precise algorithm settings are essential for accurate prediction.

Cite this article: Rezazadehvishkaei, A., Hasanibafarani, A., Pahlevanafshar, K., & Aboozari, M. (2025). Factors affecting the longevity of Holstein dairy cattle: A data mining approach. *Journal of Animal Production*, 27 (3), 255-267. DOI: <https://doi.org/10.22059/jap.2025.384401.623807>



© The Author(s).

DOI: <https://doi.org/10.22059/jap.2025.384401.623807>

Publisher: University of Tehran Press.



عوامل مؤثر بر ماندگاری گاوهای شیری هلشتاین: یک رویکرد داده کاوی

علی رضازاده ویشکایی^۱ | علیرضا حسینی بافرانی^۲ | کیان پهلوان افشار^۳ | مهران ابوذری^۴

۱. گروه علوم دامی، دانشکده کشاورزی، دانشگاه آزاد اسلامی، واحد ابهر، زنجان، ایران. رایانامه: ali.rezazadeh@iau.ac.ir
۲. نویسنده مسئول، گروه علوم دامی، سازمان تحقیقات-آموزش-ترویج کشاورزی، موسسه آموزش و ترویج کشاورزی، تهران، ایران. رایانامه: arhb@abc.org.ir
۳. گروه علوم دامی، دانشکده کشاورزی، دانشگاه آزاد اسلامی، واحد ورامین، تهران، ایران. رایانامه: pahlevanafshar.k@abhariau.ac.ir
۴. گروه علوم دامی، دانشکده کشاورزی، دانشگاه آزاد اسلامی، واحد ابهر، زنجان، ایران. رایانامه: m.aboozari1357@iau.ac.ir

اطلاعات مقاله

چکیده

نوع مقاله: مقاله پژوهشی

تاریخ دریافت: ۱۴۰۳/۰۸/۰۵

تاریخ بازنگری: ۱۴۰۴/۰۳/۱۱

تاریخ پذیرش: ۱۴۰۴/۰۳/۱۳

تاریخ انتشار: ۱۴۰۴/۰۷/۱۹

هدف: این پژوهش، به دنبال شناسایی و تجزیه و تحلیل سیستماتیک عوامل کلیدی مؤثر بر ماندگاری گاوهای شیری هلشتاین در گله، با استفاده از الگوریتم‌های پیشرفته داده کاوی است. درک و پیش‌بینی ماندگاری حیاتی است، زیرا به‌طور مستقیم بر بهره‌وری و سودآوری گاوداریهای شیری تأثیر می‌گذارد. گاوهای با طول عمر بیشتر، تمایل به تولید گوساله‌های بیشتر و تولید شیر بیشتر داشته و در نتیجه بهره‌وری اقتصادی کلی دامداری را افزایش می‌دهند. علاوه بر این، طول عمر بیشتر با کاهش هزینه‌های جایگزینی مرتبط است.

مواد و روش‌ها: در سال‌های اخیر، ادغام تکنیک‌های یادگیری ماشین در مدیریت کشاورزی و دامداری شتاب قابل‌توجهی یافته است. این مطالعه از داده‌های فنوتیپی دقیق جمع‌آوری شده از ۳۷۰۰۹ حیوان دختر متعلق به ۶۶۴ پدر در ۸۲ گله مجزا استفاده می‌کند که نشان‌دهنده یک مجموعه داده جامع است که یک دهه را دربرمی‌گیرد. داده‌ها شامل هشت رکورد تولید شیر، در کنار سایر متغیرهای مرتبط مانند سن دام، تعداد پدر، ماندگاری به ماه، تعداد سلول‌های سوماتیک، روزهای شیردهی، تولید شیر (کیلوگرم)، میزان پروتئین و چربی، طول چرخه زایش، دفعات شیردوشی، موقعیت جغرافیایی (استان)، تاریخ تولد، تاریخ زایش، فاصله زایش، کد گله و سن در اولین زایش است. مرحله آماده‌سازی داده‌ها شامل پردازش و سازماندهی مجموعه داده‌ها با استفاده از اکسل ۲۰۱۶ بود که کیفیت و سازگاری داده‌ها را تأمین نمود. تجزیه و تحلیل‌های بعدی داده‌ها با استفاده از نرم‌افزار R (نسخه ۴.۳.۳) و با استفاده از بسته‌های تخصصی مربوطه برای یادگیری ماشین و مدل‌سازی آماری انجام شد.

نتایج: نتایج نشان داد که ماشین بردار پشتیبان بهترین دقت (۰/۹۸۷) را دارد. جنگل تصادفی دومین الگوریتم کارآمد بود. دقت ماشین تقویت‌گرایان کمی کمتر از جنگل تصادفی بود، اما همچنان عملکرد خوبی را نشان داد. درخت تصمیم کم‌ترین دقت را در بین این الگوریتم‌ها ارائه داد. درخت تصمیم و ماشین بردار پشتیبان این عملکرد را با متغیرهای ورودی کم‌تری در مقایسه با ماشین تقویت‌گرایان و جنگل تصادفی به‌دست آوردند.

نتیجه‌گیری: نتایج نشان داد که هیچ یک از الگوریتم‌های مورد استفاده برای طبقه‌بندی بقاء، با وجود دقت قابل‌قبول، عاری از خطا نیستند، اما از سوی دیگر، نشان داده شد که درخت تصمیم ساده‌تر و کم‌هزینه‌تر است. مهم‌ترین ویژگی‌های این روش‌ها، عدم نیاز به فرضیات و الزامات آماری مورد نیاز روش‌های رگرسیون خطی و درونیابی، عدم نیاز به فرضیات نرمال بودن، مقاومت بودن در برابر مقادیر و مقادیر گم‌شده و توانایی تشخیص روابط غیرخطی پیچیده بین متغیرها و اهداف پیش‌بینی است که آن‌ها را برای کاربردهای مختلف در صنعت دامپروری مناسب می‌سازد. پروتکل‌های دقیق ثبت داده‌ها و همچنین تنظیمات دقیق الگوریتم برای پیش‌بینی دقیق ضروری هستند.

کلیدواژه‌ها:

ماندگاری

طول عمر گاوهای شیری

ویژگی‌های تولیدمثلی

یادگیری ماشین

الگوریتم‌های طبقه‌بندی

استناد: رضازاده ویشکایی، علی؛ حسینی بافرانی، علیرضا؛ پهلوان افشار، کیان و ابوذری، مهران (۱۴۰۴). عوامل مؤثر بر ماندگاری گاوهای شیری هلشتاین: یک رویکرد داده کاوی. نشریه توليدات دامی، ۲۷ (۳)، ۲۶۷-۲۵۵. DOI: <https://doi.org/10.22059/jap.2025.384401.623807>

ناشر: مؤسسه انتشارات دانشگاه تهران.

© نویسندگان.



۱. مقدمه

یادگیری ماشین در مطالعات علوم دامی برای پیش‌بینی صفات اقتصادی مورد استفاده قرار گرفته است (Shahinfar et al., 2021). مزیت کلیدی این روش‌ها، توانایی آن‌ها در استخراج روابط بین متغیرها در داده‌های بزرگ و پیچیده است. بسیاری از مدل‌های پارامتری به فرضیات خاصی درباره توزیع داده‌ها و شکل رابطه‌های بین متغیرها نیاز دارند (مانند نرمال بودن، خطی بودن و غیره)، درحالی‌که روش‌های یادگیری ماشین می‌توانند این فرضیات را نادیده بگیرند و از داده‌ها به صورت تجربی یاد بگیرند. همچنین بسیاری از روش‌های یادگیری ماشین مانند درخت تصمیم و ماشین‌های بردار پشتیبان می‌توانند روابط غیرخطی و پیچیده بین متغیرها را به خوبی شناسایی کنند، درحالی‌که مدل‌های پارامتری معمولاً فقط قادر به مدل‌سازی روابط خطی هستند (Hastie et al., 2009). کاربرد ساده‌تر و انعطاف‌پذیری این روش‌ها، استفاده از آن‌ها را در پژوهش‌های گسترش داده است (Carvalho et al., 2019). پژوهش‌گران معمولاً، مدل‌های یادگیری ماشین را برای یافتن مدلی که بتواند دقت پیش‌بینی را برای یک پدیده خاص بهینه کند، بررسی می‌کنند (Curtis et al., 2019). از الگوریتم‌های یادگیری ماشین به طور گسترده برای حل مسائل مختلف طبقه‌بندی و پیش‌بینی در گاوهای شیری استفاده شده است. الگوریتم‌هایی نظیر ماشین‌های بردار پشتیبان، یا جنگل تصادفی، با دقت ۸۵ تا ۹۰ درصد برای طبقه‌بندی ورم پستان در گاوهای شیری استفاده شده‌اند (Fadul-Pacheco et al., 2021). ماندگاری فاصله زمانی بین زایش اول تا زمان حذف حیوان ماده از گله است (Hu et al., 2021). عوامل متعددی نظیر خصوصیات تیپی، عملکرد تولید مثلی و تولیدی بر طول مدت ماندگاری گاو در گله تأثیر دارند. صفات تولیدمثلی از مهم‌ترین عوامل مؤثر بر ماندگاری هستند که در تعیین کارایی اقتصادی سیستم‌های تولید گاو اهمیت دارند. کاهش ماندگاری گاو، همواره برای صنعت گاو شیری نگران‌کننده بوده است (Schuster et al., 2020). سن در اولین زایش، مرحله شیردهی، تولید شیر، تنوع در اندازه گله، و همچنین مشکلات باروری، سلامت پستان و بیماری‌ها عواملی هستند که به طور قابل توجهی در حذف زود هنگام گاو از گله نقش دارند. محتوای چربی شیر، شمارش سلول‌های بدنی (SCC) و غلظت بتا هیدروکسی بوتیرات (BHB) در شیر نیز در ماندگاری مؤثرند (Rostellato et al., 2022). ماندگاری گاوهای شیری در گله، به طور متوسط حدود ۵/۸ سال است که از ۴/۹ سال تا ۷/۱ سال متغیر بوده و کم‌تر از طول عمر طبیعی گاوهای شیری است (Van Pelt et al., 2017). اهمیت ماندگاری در گله گاوهای شیری را می‌توان از این منظر مورد توجه دانست که به طور معمول تولید سالانه در دوره پنجم شیردهی به حداکثر می‌رسد و معمولاً بیش‌ترین سود سالانه در دوره ششم شیردهی به دست می‌آید (Horn et al., 2013). برخی از پژوهش‌گران نشان دادند که بالاترین عملکرد شیردهی پس از زایمان سوم رخ می‌دهد، بنابراین نگهداری حیوان برای مدت طولانی‌تر منجر به افزایش تولید می‌شود (Elahi Torshizi et al., 2019). به همین دلیل در کشورهای مختلف ماندگاری به عنوان یکی از صفات مهم اقتصادی در شاخص انتخاب گاوهای نر شیری منظور شده است (Imbayarwo-Chikosi et al., 2017). عموماً وراثت‌پذیری این صفت کم‌تر از ۱۰ درصد محاسبه می‌شود که نشان می‌دهد این صفت بیش‌تر تحت تأثیر عوامل محیطی است (Hu et al., 2021). هدف پژوهش حاضر بررسی کارایی روش‌های منتخب یادگیری ماشین شامل جنگل تصادفی (RF)، گرادیان بوستینگ ماشین (GBM)، درخت تصمیم (DT) و ماشین بردار پشتیبان (SVM)، در پیش‌بینی عوامل مؤثر بر طول مدت ماندگاری گاوهای هلستاین ایران در گله بوده است.

۲. روش پژوهش

۲.۱. مجموعه داده‌ها

داده‌های مورد استفاده در این پژوهش به وسیله مرکز اصلاح نژاد دام و بهبود تولیدات دامی (وزارت جهاد کشاورزی) تأمین شد. در این مطالعه از داده‌های فنوتیپی مربوطه به ۳۷۰۰۹ دختر از ۶۶۴ پدر در ۸۲ گله از یازده استان، که دارای

هشت رکورد تولید شیر مربوط به ۱۰ سال اخیر منتهی به ۱۴۰۳ بودند استفاده شد (جدول ۱). طول عمر حیوانات به عنوان دوره بین تاریخ تولد و آخرین تاریخ ثبت تعریف و براساس ماه محاسبه شد. در هنگام استفاده از تکنیک‌های طبقه‌بندی می‌توان متغیرهای پیوسته‌ای نظیر ماندگاری را در قالب کلاس، به متغیر گسسته تبدیل کرد. این کار اجازه می‌دهد تا الگوها، روندها و روابط بین سایر متغیرهای و متغیر مورد نظر آسان‌تر شناسایی شود. تبدیل متغیر پیوسته به متغیر گسسته دلایل کاربردی مختلفی دارد که برخی از آن‌ها عبارتند از سادگی و تفسیرپذیری، کاهش پیچیدگی داده‌ها، بهبود کارایی مدل‌ها و کنترل بر خاصیت‌های ناهمگنی متغیرها. به همین منظور در این پژوهش، داده‌های ماندگاری به صورت کلاس‌بندی به کار گرفته شد. نحوه تعریف کلاس‌ها به این صورت بود که اگر مدت ماندگاری کمتر از ۱۹ ماه بود کلاس E، بین ۱۹ تا ۳۹ ماه کلاس D، بین ۳۹ تا ۵۹ ماه کلاس C، بین ۵۹ تا ۷۹ ماه کلاس B و بیش از ۷۹ ماه کلاس A در نظر گرفته شد. فایل اطلاعات شامل متغیرهای متنوعی نظیر شماره حیوان، سن، شماره پدر، مدت ماندگاری به ماه، شمارش سلول‌های بدنی، روزهای شیردهی، مقدار تولید شیر، پروتئین و چربی شیر (کیلوگرم)، نوبت زایش، دفعات دوشش، استان، تاریخ تولد، تاریخ زایش، فاصله زایش، کد گله و سن در اولین زایش و غیره بود. در مرحله آماده‌سازی داده‌ها از نرم‌افزار صفحه گسترده اکسل ۲۰۱۶ و برای تجزیه و تحلیل آماری از نرم‌افزار آماری R (نسخه ۴,۳,۳) و بسته‌های مربوطه استفاده شد.

جدول ۱. آمار توصیفی اطلاعات مورد استفاده

اطلاعات	تعداد حیوانات
تعداد حیوانات رد	۳۷۰۰۹
تعداد پدران	۶۶۴
مولد ماده دارای نتاج	۲۵۳۱۶
تعداد گله‌ها (از ۱۱ استان کشور)	۸۲

۲.۲. تحلیل داده‌ها

آماده‌سازی داده‌ها با استفاده از اکسل ۲۰۱۶ انجام شد. تجزیه و تحلیل داده‌ها با استفاده از نرم‌افزار R (نسخه ۴,۳,۳) و بسته‌های مربوطه انجام شد. الگوریتم‌های مورد استفاده شامل جنگل تصادفی (Random Forest; RF)، ماشین تقویت گرادیان (Gradient Boosting Machines; GBM)، درخت تصمیم (Decision Tree; DT) و ماشین بردار پشتیبان (Support Vector Machine; SVM) بودند که در ادامه به توصیف هر یک پرداخته شده است.

۳.۲. جنگل تصادفی

متداول‌ترین پکیج برای پیاده‌سازی مدل‌های جنگل تصادفی در نرم‌افزار R، randomForest است. بسته caret نیز طیف گسترده‌ای از ابزارها را برای مدل‌های آموزشی ارائه می‌دهد و به پیش پردازش و ارزیابی مدل کمک می‌کند. بسته Ranger برای پیاده‌سازی سریع جنگل تصادفی، به‌ویژه در مجموعه داده‌های بزرگ مفید است و عملکرد خوبی را برای وظایف طبقه‌بندی و رگرسیون ارائه می‌دهد (Kuhn., 2019). پس از آماده‌سازی، داده‌ها به‌طور تصادفی به ۷۰ درصد برای آموزش و ۳۰ درصد برای مجموعه داده‌های آزمایشی تقسیم شدند. در مجموعه داده آموزشی، از تابع trainControl برای انتخاب هایپرپارامترها و اعتبارسنجی مدل شد. این تابع به‌طور خاص، می‌تواند نحوه تقسیم و ارزیابی داده‌های آموزشی را مشخص کرده و اغلب از روش اعتبارسنجی متقاطع k-fold برای ارزیابی بهتر مدل استفاده می‌کند. برای این منظور مجموعه داده به پنج فولد تقسیم شد. به‌منظور بهینه‌سازی هایپرپارامترها، تکنیک‌هایی مانند جستجوی شبکه‌ای،

جستجوی تصادفی، بهینه‌سازی بیزی، الگوریتم‌های ژنتیک و بهینه‌سازی مبتنی بر گرادیان را می‌توان برای یافتن مقادیر بهینه برای فرایمترها به کار برد (Bergstra & Yoshua, 2012). در مطالعه حاضر، از مقادیر پیش‌فرض و از طریق فرایند جستجوی تصادفی از طریق بسته caret بهره گرفته شد. مدل با مجموعه پارامترهایی که بالاترین دقت را داشت برای پیش‌بینی استفاده شد. در این پژوهش مقدار پیش‌فرض mtry (تعداد ویژگی‌ها در هر تقسیم) برابر ۱۹ و مقدار پیش‌فرض تعداد درختان $n_{tree}=500$ در نظر گرفته شد. اهمیت ورودی‌ها با درصد افزایش میانگین مربعات خطای پیش‌بینی ($IncMSE\%$) شناسایی می‌شود.

۴.۲. گرادیان بوستینگ ماشین

در روش گرادیان بوستینگ ماشین، اهمیت ورودی با تأثیر نسبی آن بر کاهش میانگین مربعات خطا در تمام درخت‌ها تعیین می‌شود. هر متغیری که بتواند خطای مدل را کاهش دهد، به‌عنوان متغیر مهم انتخاب شده و در گره قرار می‌گیرد. ریشه میانگین مربعات خطا (RMSE) و میانگین خطای مطلق (MAE) براساس مقدار اندازه‌گیری شده برآورد شد، میانگین خطای مطلق در مقایسه با ریشه میانگین مربعات خطا کم‌تر آسیب‌پذیر است (Botchkarev, 2018). مقادیر پایین‌تر RMSE و MAE با نتایج پیش‌بینی بهتر همراه است. هم‌چنین ضریب تعیین (R^2) که بیانگر میزان تغییرات توضیح داده‌شده توسط مدل است، نیز برآورد شد. از تابع ماتریس درهم ریختگی (confusionMatrix) برای ارزیابی عملکرد مدل برای هر مجموعه داده آموزشی - آزمایشی استفاده شد.

۵.۲. درخت تصمیم

این الگوریتم درخت‌های باینری ایجاد می‌کند، به‌طوری‌که از هر گره داخلی دو لبه خارج می‌شود و درخت‌های به‌دست‌آمده توسط روش اثربخشی هزینه، هرس می‌شوند. برای هر نتیجه آزمایش ممکن در گره، یک شاخه ساخته می‌شود که به یک گره دختر ختم می‌شود. الگوریتم برای تفکیک، میزان حداقل مربع خطا را جستجو می‌کند (Witten & Frank., 2005). در یک مطالعه نشان داده شد که استفاده از الگوریتم درخت تصمیم برای طبقه‌بندی دقیق سطح فعالیت خوردن، نشخوارکردن و خوابیدن امکان‌پذیر است (Tamura *et al.*, 2021). پیترزما و همکاران، درخت تصمیم را برای مجموعه داده‌های کوچک نیز مفید دانسته‌اند (Pietersma *et al.*, 2003). رابرت و همکاران، الگوریتم درخت تصمیم را برای طبقه‌بندی رفتارهای مختلف در گاو پیاده‌سازی کردند. بیش‌ترین دقت آزمایش آن‌ها $84/10$ درصد و میانگین دقت طبقه‌بندی $82/72$ درصد بود (Robert *et al.*, 2009). برای برقراری درخت‌های رگرسیون و دسته‌بندی از الگوریتم CART استفاده شد.

۶.۲. ماشین بردار پشتیبان

ماشین بردار پشتیبان دو پارامتر مهم دارد. الف- پارامتر C که طبقه‌بندی‌های نادرست را جریمه می‌کند. به این صورت که مقدار C کم منجر به یک مدل ساده با حاشیه نرم می‌شود، و مدل با مقدار C بزرگ، طبقه‌بندی کامل را در اولویت قرار می‌دهد؛ ب- پارامتر γ که نقش یک پیکسل آموزشی منفرد را مشخص می‌کند. مقادیر بسیار کوچک آن منجر به مدل‌های محدود و مقادیر بسیار بزرگ منجر به بیش از حد برازش می‌شوند (Lantz., 2013). از تکنیک‌های ماشین بردار پشتیبان برای تعیین وجود ورم پستان تحت بالینی و بالینی در گاوهای شیری براساس شمارش سلول‌های سوماتیک سالم یا آلوده با حساسیت ۸۹ درصد، ویژگی ۹۲ درصد استفاده شده است (Mammadova & Keskin., 2013). این تکنیک داده‌های ورودی را با تولید یک ابر صفحه و به‌صورت بهینه، به کلاس‌های مجزا متمایز می‌کند و این کار را

براساس اصل به حداقل رساندن ریسک ساختاری انجام می‌دهد. به عبارت دیگر، هدف یافتن حالت حداکثر فاصله بین دو کلاس است. ماشین بردار پشتیبان در مسائل طبقه‌بندی مانند تشخیص الگو، تشخیص متن و طبقه‌بندی پروتئین استفاده شده است و نتایج با کیفیت بالایی را به همراه داشته است (Smola & Schölkopf., 2004). در این پژوهش مقادیر نهایی مورد استفاده γ برابر با $0/001$ و مقدار C برابر با ۱۰۰۰ بود.

۷.۲. تیون کردن و انتخاب بهترین هایپرپارامتر

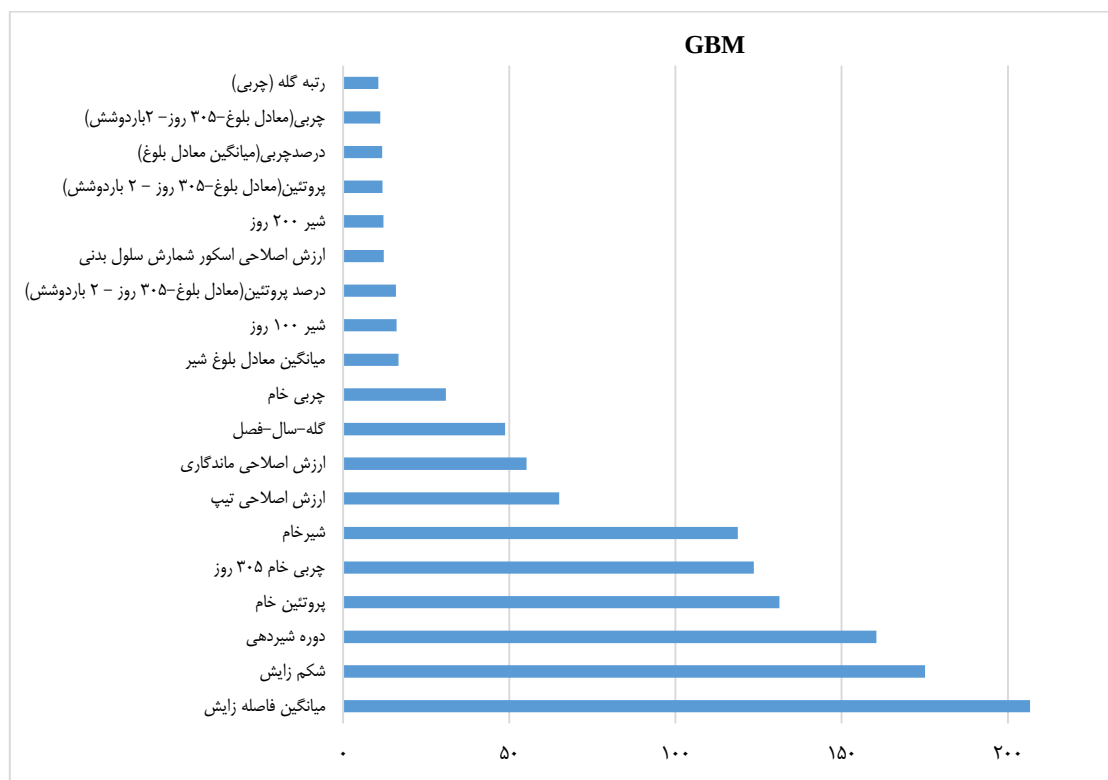
برای تیون کردن و انتخاب بهترین هایپرپارامتر در روش گرادیان بوستینگ ماشین از کل داده‌ها ۷۰ درصد نمونه‌گیری کردیم و با روش Cross-Validated و در پنج فولد، بهترین هایپر پارامتر مشخص شد. مقادیر نهایی مورد استفاده برای مدل با در نظر گرفتن بیش‌ترین دقت (Accuracy)، به صورت $n.trees=3000$ ، $interaction.depth=1$ تعیین شد. مقدار Shrinkage که سرعت یادگیری مدل را کنترل می‌کند و به محدود کردن بیش از حد برازش کمک می‌کند برابر با $0/2$ و $n.minobsinnode$ که حداقل تعداد مشاهدات در گره‌های پایانی را مشخص می‌کند برابر با ۵۰ در نظر گرفته شد. برای تیون کردن و انتخاب بهترین هایپر پارامتر در روش RF به تریبی مشابه، بهترین هایپر پارامتر مشخص شد. مقادیر نهایی مورد استفاده برای مدل $mtry$ برابر با ۲۸، $ntree$ برابر با ۴۰۰۰ و $nodesize$ برابر با ۱۰۰ محاسبه شد. برای تیون کردن و انتخاب بهترین هایپر پارامتر در روش درخت تصمیم از کل داده‌ها ۷۰ درصد نمونه‌گیری کردیم و با روش Cross-Validated، بهترین هایپر پارامتر مشخص شد. برای انتخاب مدل بهینه از بالاترین Accuracy استفاده شد. پارامتر پیچیدگی (cp) یک پارامتر توقف است که به سرعت بخشیدن به جستجوی تقسیم‌ها کمک می‌کند و می‌تواند شکاف‌هایی که این معیار را برآورده نمی‌کنند شناسایی و آن‌ها را هرس کند. مقدار نهایی برای آن $cp=0/0023$ محاسبه شد. کاپا کوهن اساساً با توجه به فرکانس هر کلاس به ما می‌گوید که طبقه‌بندی مورد نظر، در مقایسه با شرایط طبقه‌بندی تصادفی، چقدر بهتر عمل می‌کند. کاپا کوهن همیشه کمتر یا مساوی ۱ است. مقادیر صفر یا کمتر نشان می‌دهد که طبقه‌بندی کننده بی فایده است. برای تیون کردن و انتخاب بهترین هایپر پارامتر در روش ماشین بردار پشتیبان نیز از کل داده‌ها ۷۰ درصد نمونه‌گیری کردیم و با روش Cross-Validated و در پنج فولد، بهترین هایپر پارامتر مشخص شد. مقادیر نهایی مورد استفاده (γ) برای مدل برابر با $0/001$ و مقدار C برابر با ۱۰۰۰ بود.

۳. یافته‌های پژوهش

۳.۱. گرادیان بوستینگ ماشین

مؤثرترین متغیرها برای پیش‌بینی ماندگاری در این الگوریتم، براساس اندازه‌گیری تأثیر نسبی، در نمودار شکل (۱) ارائه شده‌اند.

همان‌طور که مشاهده می‌شود متغیرهای متوسط فاصله زایش، تعداد شکم زایش و دوره شیردهی بیش‌ترین تأثیر نسبی را در پیش‌بینی داشته‌اند که البته به دلیل همبستگی بالا با ماندگاری قابل انتظار است. اما در مرحله بعد متغیرهایی تولیدی نظیر پروتئین خام، چربی خام، شیرخام، معادل بلوغ شیر و چربی و پروتئین تولیدی، شیر ۱۰۰ و ۲۰۰ روز و البته ارزش اصلاحی تیپ و شمارش سلول‌های بدنی شیر و در نهایت رتبه‌بندی گله براساس چربی شیر، تأثیر خوبی را بر پیش‌بینی داشته‌اند، به طوری که سرجمع تأثیر این متغیرها، نسبت به سه متغیر اول، به گونه‌ای است که به خوبی امکان پیش‌بینی زود هنگام ماندگاری را به دست می‌دهند. مدل مورد استفاده در این الگوریتم توانست در هر پنج تکرار مطابقت حدود ۹۳ درصد را در پیش‌بینی کلاس‌ها نشان دهد.



شکل ۱. متغیرهای مؤثر (تأثیر نسبی) برای پیش‌بینی ماندگاری در این الگوریتم گرادینت بوستینگ ماشین

۲.۳. جنگل تصادفی

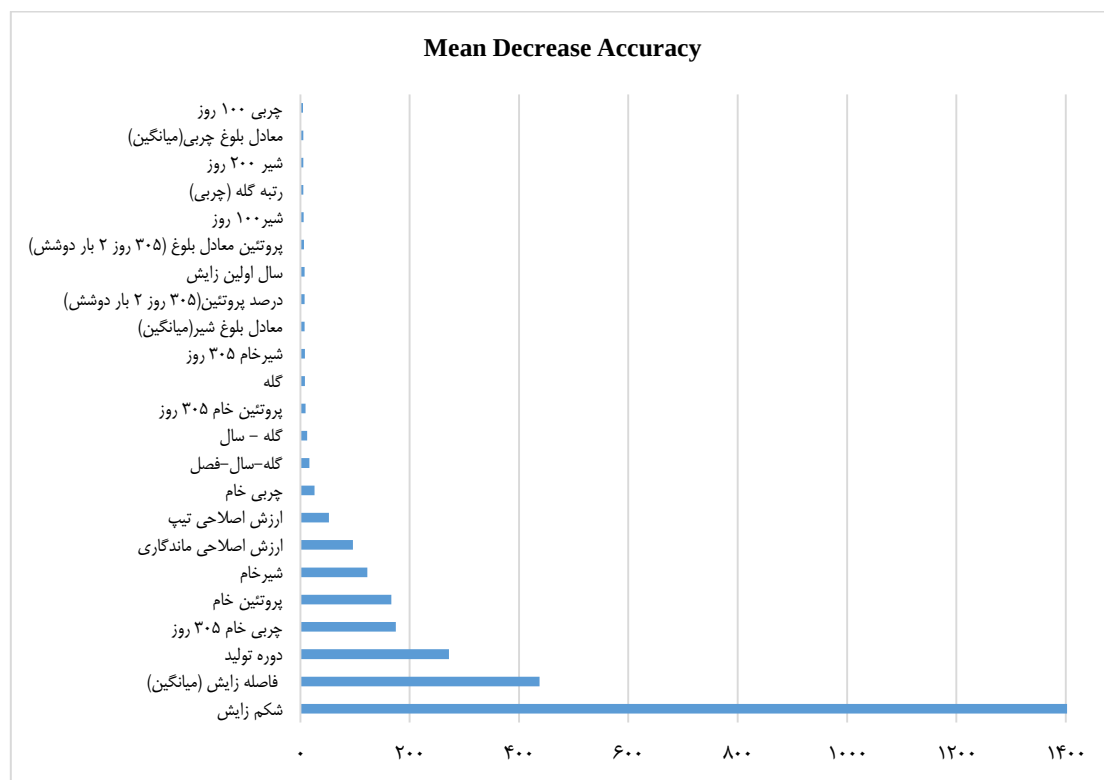
برای بررسی تأثیر متغیرهای در قابلیت پیش‌بینی مدل از نمودار میانگین کاهش دقت (Mean Decrease Accuracy) و میانگین کاهش ضریب جینی (Mean Decrease Gini) استفاده شد. هرچه مقدار میانگین کاهش دقت یا میانگین کاهش نمره جینی بیش‌تر باشد، اهمیت متغیر در مدل بالاتر است. مؤثرترین متغیرها برای پیش‌بینی ماندگاری براساس میانگین کاهش دقت، در شکل (۲) و براساس کاهش نمره جینی در شکل (۳) نمایش داده شده‌اند.

براساس میانگین تأثیر بر کاهش دقت مدل، مؤثرترین متغیرها شکم زایش، متوسط فاصله زایش، چربی خام، پروتئین و شیر خام و دوره شیرواری بودند. در مرحله دوم اهمیت متغیرهایی نظیر ارزش اصلاحی تیپ، گله، تولید شیر و چربی ۳۰۵ روز و معادل بلوغ آن‌ها، چربی ۱۰۰ و شیر ۲۰۰ روز نقش آفرین بوده‌اند.

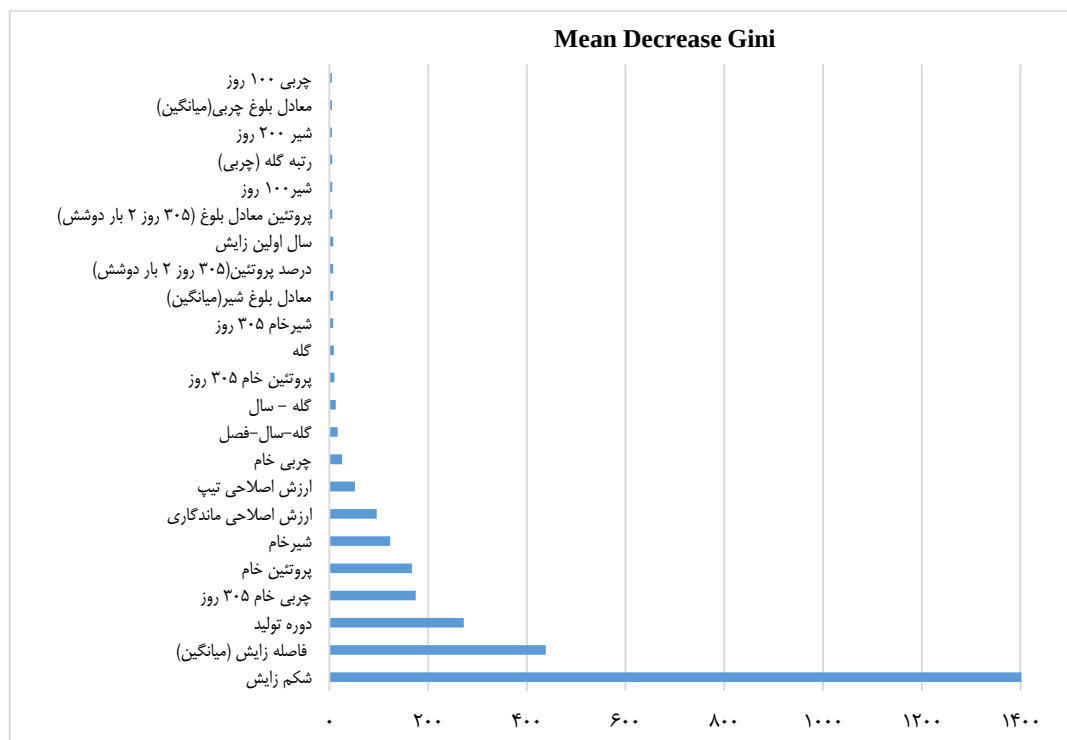
متوسط کاهش ضریب جینی، متغیرهای شکم زایش، فاصله زایش و دوره شیرواری را مؤثرترین متغیرهای نشان داد و در مرحله دوم تولید شیر، چربی و پروتئین خام، ارزش اصلاحی ماندگاری و ارزش اصلاحی تیپ حیوان را مؤثرترین متغیرهای نشان داد. نرخ مطابقت کلاس‌های پیش‌بینی شده با مقادیر مشاهده‌شده، در هر پنج تکرار این روش حدود ۹۳ درصد محاسبه شد.

۳.۳. روش درخت تصمیم

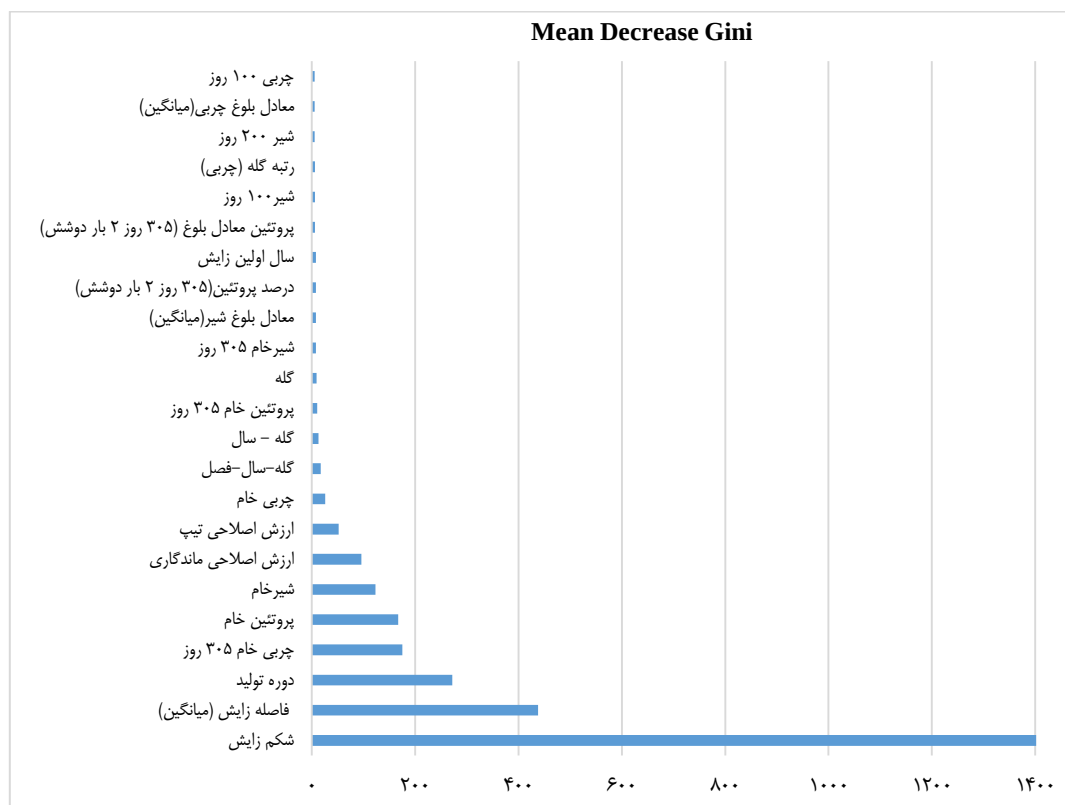
متغیرهایی که بیش‌ترین تأثیر (Overall) را در قابلیت پیش‌بینی کلاس‌ها توسط این الگوریتم داشته‌اند، در نمودار زیر نمایش داده شده‌اند.



شکل ۲. متغیرهای مؤثر (تأثیر نسبی) برای پیش‌بینی ماندگاری در الگوریتم جنگل تصادفی براساس میانگین کاهش دقت



شکل ۳. متغیرهای مؤثر (تأثیر نسبی) برای پیش‌بینی ماندگاری در الگوریتم جنگل تصادفی براساس کاهش نمره جینی



شکل ۴. مؤثرترین متغیرها (تأثیر نسبی) برای پیش‌بینی ماندگاری در الگوریتم درخت تصمیم

همان‌طور که از نمودار فوق قابل مشاهده است شکم زایش، فاصله زایش، دوره تولید، تولید شیر خام، تولید چربی خام، تولید پروتئین خام، ارزش اصلاحی ماندگاری و ارزش اصلاحی تیپ بیش‌ترین تأثیر را در پیش‌بینی الگوریتم داشته‌اند. نرخ مطابقت پیش‌بینی کلاس‌های با مقادیر مشاهده شده نیز در هر پنج تکرار حدود ۹۳ درصد محاسبه شد.

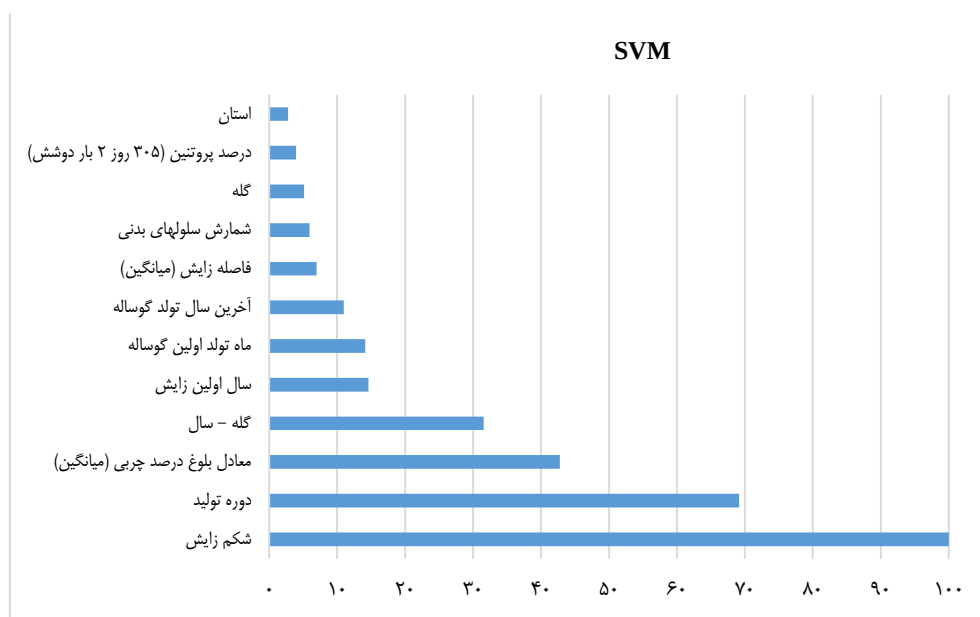
۴.۳. ماشین بردار پشتیبان

متغیرهایی که بیش‌ترین تأثیر را در قابلیت پیش‌بینی کلاس‌ها توسط این الگوریتم داشته‌اند، در نمودار زیر نمایش داده شده‌اند.

همان‌طور که از نمودار فوق قابل مشاهده است متغیرهای شکم زایش، دوره تولید، میانگین معدل بلوغ چربی شیر، بیش‌ترین تأثیر و در مرحله بعد، شمارش سلول‌های بدنی شیر، گل، مقدار پروتئین ۳۰۵ روز/ دوبار دوشش و استان مربوطه در نتایج مؤثر بوده‌اند. نرخ مطابقت پیش‌بینی کلاس‌های با مقادیر مشاهده شده، در این الگوریتم در هر پنج تکرار حدود ۹۳ درصد محاسبه شد.

۴.۴. بحث

به‌منظور مقایسه عملکرد الگوریتم‌های مورد استفاده در پیش‌بینی کلاس ماندگاری، از آزمون t و آنالیز واریانس استفاده شد. نتایج عملکرد کلی الگوریتم‌ها در پنج فولد در جدول (۲) قابل مشاهده است.



شکل ۵. مؤثرترین متغیرها (درصد) برای پیش‌بینی ماندگاری در الگوریتم ماشین بردار پشتیبان

جدول ۲. عملکرد کلی الگوریتم‌های مورد استفاده در پنج فولد

cc	GBM	RF	DT	SVM
fold1	۰٫۹۷۶۹۷۹	۰٫۹۸۳۵۹۶	۰٫۹۳۲۲۸۳	۰٫۹۸۸۹۵۹
fold2	۰٫۹۸۰۳۱۱	۰٫۹۸۳۳۱۴	۰٫۹۲۳۶۶۵	۰٫۹۸۸۳۸۲
fold3	۰٫۹۷۹۴۳۲	۰٫۹۸۲۶۴۱	۰٫۹۲۶۵۷۵	۰٫۹۸۷۴۶۲
fold4	۰٫۹۸۱۹۳۹	۰٫۹۸۴۲۳۶	۰٫۹۲۴۹۴	۰٫۹۸۹۲۷۱
fold5	۰٫۹۷۷۱۷۹	۰٫۹۸۳۲۹۷	۰٫۹۱۶۵۵۲	۰٫۹۸۸۷۹۲

داده‌های جدول فوق نشان می‌دهد که به‌طور کلی الگوریتم ماشین بردار پشتیبان (Support Vector Machine; SVM) بهترین دقت را در تمام پنج فولد دارد، و مقادیر دقت بالای ۰/۹۸۷ نشان‌دهنده کارایی بالای این الگوریتم است. الگوریتم جنگل تصادفی (Random Forest; RF) دومین الگوریتم کارآمد است. دقت الگوریتم ماشین تقویت گرادیان (Gradient Boosting Machines; GBM)، نسبت به جنگل تصادفی کمی کمتر است، اما هنوز هم عملکرد مناسبی دارد. الگوریتم درخت تصمیم (Decision Tree; DT) کم‌ترین دقت را در میان این الگوریتم‌ها ارائه می‌دهد. هم‌چنین می‌توان دید که تمامی الگوریتم‌ها، طی پنج فولد دارای نوسانات مختصری در دقت هستند. مقایسه فولدها نشان می‌دهد که در fold1، الگوریتم ماشین بردار پشتیبان بهترین عملکرد را دارد، درحالی‌که درخت تصادفی به‌وضوح کم‌تر از بقیه است. هم‌چنین می‌توان دید که در fold4، الگوریتم جنگل تصادفی و ماشین بردار پشتیبان هر دو عملکرد بسیار خوبی دارند، درحالی‌که درخت تصمیم هنوز هم به‌طور نسبی پایین‌تر است. به‌منظور بررسی اینکه آیا الگوریتم‌های یادگیری ماشین مورد استفاده، با یکدیگر تفاوت معنی‌داری دارند یا خیر. از تحلیل واریانس (ANOVA) استفاده شد که نتایج آن در جدول (۳) زیر ارائه شده است.

جدول ۳. جدول تجزیه واریانس عملکرد الگوریتم‌های استفاده‌شده در پنج فولد

منابع واریانس	SS	df	MS	F	P-value	F بحرانی
بین گروه	۰/۰۰۰۷۵۳	۳	۰/۰۰۰۲۵۱	۱۳/۲۷۳۴۲	۰/۰۰۰۱۳۱	۳٫۲۲۸۸۷۲
داخل گروه	۰/۰۰۰۳۰۳	۱۶	۰/۰۰۰۰۱۸۹			
کل	۰/۰۰۱۰۵۶	۱۹				

همان طور که از جدول فوق قابل مشاهده است، مقدار F محاسباتی (۱۳/۲۷) به طور قابل توجهی بزرگتر از مقدار بحرانی (۳/۲۳) است که نشان می‌دهد تفاوت بین میانگین‌های الگوریتم‌ها معنی‌دار است. مقدار P-value بسیار کمتر (۰/۰۰۰۱۳۱) از سطح اهمیت ۰/۰۵ است که بیانگر تفاوت معنی‌دار حداقل یک گروه با دیگر گروه‌هاست. برای بررسی این که کدام گروه‌ها تفاوت دارند، از آزمون t زوجی استفاده شد. نتایج در جدول (۴) ارائه شده است.

جدول ۴. مقایسه عملکرد الگوریتم‌های مورد استفاده در پیش‌بینی کلاس ماندگاری، با استفاده از آزمون زوجی t

مقایسه زوجی الگوریتم‌ها	میانگین	انحراف استاندارد	میانگین خطای استاندارد	t	df	Sig. (2-tailed)
GBM - RF	۰/۰۰۴۲	۰/۰۰۴۱	۰/۰۰۱۸	۲/۳۰۶	۴	۰/۰۸۲
GBM - DT	۰/۰۱۳۹	۰/۰۰۵۹	۰/۰۰۲۷	۵/۲۳۹	۴	۰/۰۰۶
GBM - SVM	-۰/۰۰۱۹	۰/۰۰۳۹	۰/۰۰۱۸	-۱۰/۱۰۹	۴	۰/۳۳
RF - DT	۰/۰۰۹۷	۰/۰۰۶۲	۰/۰۰۲۸	۳/۵	۴	۰/۰۲۵
RF - SVM	۰/۰۰۶۲	۰/۰۰۴۷	۰/۰۰۲۱	-۲/۹۰۷	۴	۰/۰۴۴
SVM - DT	-۰/۰۱۶	۰/۰۰۵۳	۰/۰۰۲۴	-۶/۷۱	۴	۰/۰۰۳

از داده‌های جدول فوق می‌توان دید بین عملکرد الگوریتم ماشین گرادیان تقویت شده و جنگل تصادفی (GBM-RF)، و عملکرد ماشین گرادیان تقویت شده و ماشین بردار پشتیبان (GBM-SVM) تفاوت معنی‌دار وجود ندارد. اما سایر مقایسات دو دویی تفاوت‌های معنی‌دار را نشان می‌دهند، به طوری که بین عملکرد ماشین گرادیان تقویت شده و ماشین بردار پشتیبان با الگوریتم درخت تصمیم بیش‌ترین تفاوت مشاهده می‌شود. با بررسی این جدول، می‌توان نتیجه‌گیری کرد که الگوریتم‌های GBM و DT بهترین عملکرد را از خود نشان داده‌اند و تفاوت‌های آن‌ها با سایر الگوریتم‌ها به طور معنی‌داری بیش‌تر است. در واقع با تلفیق نتایج دو جدول (۳) و (۴) فوق، به نظر می‌رسد که الگوریتم ماشین بردار پشتیبان بالاترین دقت را داشته باشد. جدول (۵) مؤثرترین متغیرهای به کار گرفته شده توسط هر الگوریتم را نشان می‌دهد. به منظور سهولت نشان دادن رابطه بین الگوریتم‌ها در به کارگیری متغیرها، متغیرهای مشترک بین الگوریتم‌ها، به کمک رنگ‌های متمایز دسته‌بندی شده‌اند.

جدول ۵. متغیرهای به کار گرفته شده (از میان ۵۷ متغیر ورودی) توسط هر الگوریتم

DT	GBM	RF	SVM
میانگین فاصله زایش	میانگین فاصله زایش	میانگین فاصله زایش	میانگین فاصله زایش
چربی خام ۳۰۵ روز	چربی خام ۳۰۵ روز	چربی خام ۳۰۵ روز	چربی خام ۳۰۵ روز
گله - سال	گله - سال	گله - سال	گله - سال
شکم زایش	شکم زایش	شکم زایش	شکم زایش
دوره شیرواری	دوره شیرواری	دوره شیرواری	دوره شیرواری
شیرخام	شیرخام	شیرخام	شیرخام
	پروتئین (۳۰۵ روز ۲/ بار دوشش)	پروتئین (۳۰۵ روز ۲/ بار دوشش)	پروتئین (۳۰۵ روز ۲/ بار دوشش)
	شیرخام ۳۰۵ روز	شیرخام ۳۰۵ روز	شیرخام ۳۰۵ روز
	رتبه گله در پروتئین	رتبه گله در پروتئین	رتبه گله در پروتئین
	استان	استان	استان
	شیر ۱۰۰ روز	شیر ۱۰۰ روز	شیر ۱۰۰ روز
	شیر (۳۰۵ روز ۲/ بار دوشش)	شیر (۳۰۵ روز ۲/ بار دوشش)	شیر (۳۰۵ روز ۲/ بار دوشش)
	چربی (۳۰۵ روز ۲/ بار دوشش)	چربی (۳۰۵ روز ۲/ بار دوشش)	چربی (۳۰۵ روز ۲/ بار دوشش)
	درصد چربی (میانگین معادل بلوغ)	درصد چربی (میانگین معادل بلوغ)	درصد چربی (میانگین معادل بلوغ)
پروتئین خام	پروتئین خام	پروتئین خام	
پروتئین خام ۳۰۵ روز	پروتئین خام ۳۰۵ روز	پروتئین خام ۳۰۵ روز	
گله - سال - فصل	گله - سال - فصل	گله - سال - فصل	
ارزش اصلاحی ماندگاری	ارزش اصلاحی ماندگاری	ارزش اصلاحی ماندگاری	
چربی خام	چربی خام	چربی خام	
ارزش اصلاحی تیپ	ارزش اصلاحی تیپ	ارزش اصلاحی تیپ	
گله		گله	گله
سال اولین زایش		سال اولین زایش	سال اولین زایش

همان‌طور که مشاهده می‌شود الگوریتم درخت تصمیم و الگوریتم ماشین بردار پشتیبان، با تعداد متغیرهای کم‌تر توانسته‌اند با دقتی مشابه (۹۳ درصد) با سایر الگوریتم‌ها، کلاس ماندگاری را پیش‌بینی نمایند. می‌توان دید که الگوریتم‌های درخت تصمیم، ماشین گرادیان بوستینگ، درخت تصادفی و ماشین بردار پشتیبان به ترتیب از ۱۴، ۲۰، ۲۲ و ۱۶ متغیر برای حصول دقت مشابه (۹۳ درصد) استفاده کرده‌اند.

۵. نتیجه‌گیری و پیشنهادها

نتایج یافته‌ها نشان می‌دهد که هیچ یک از الگوریتم‌های بکار رفته برای طبقه‌بندی ماندگاری، با وجود دقت قابل قبول عاری از خطا نیستند، اما از طرف دیگر نشان داده شد که الگوریتم درخت تصمیم ساده‌تر و کم‌هزینه‌تر عمل می‌کند. مهم‌ترین ویژگی این روش‌ها عدم نیاز به فرضیات آماری و الزاماتی که رگرسیون خطی و روش‌های درون‌یابی نیاز دارند، عدم نیاز به مفروضات نرمال بودن، استحکام نسبت به مقادیر و مقادیر از دست‌رفته و توانایی تشخیص روابط پیچیده غیرخطی بین متغیرها و اهداف پیش‌بینی است که آن‌ها را برای کاربردهای مختلف صنعت دام مناسب می‌سازد. پروتکل‌های دقیق ثبت داده‌ها و همچنین تنظیمات دقیق الگوریتم‌ها برای پیش‌بینی دقیق ضروری است.

۶. ملاحظات اخلاقی

نویسندگان اصول اخلاقی را در انجام و انتشار این پژوهش علمی رعایت نموده‌اند و این موضوع مورد تأیید همه آن‌هاست.

۷. مشارکت نویسندگان

علی رضا زاده ویشکایی: تهیه و آماده‌سازی داده‌ها، تجزیه و تحلیل داده‌ها، تهیه پیش‌نویس مقاله؛
علیرضا حسنی بافرانی: تجزیه و تحلیل آماری داده‌ها، همکاری در نگارش، بازبینی و نهایی‌سازی مقاله؛
کیان پهلوان افشار: بررسی روش کار، بازبینی نتایج و همکاری در نگارش مقاله؛
مهران ابوذری: بازبینی نتایج و همکاری در نگارش مقاله.

۸. تعارض منافع

هیچ‌گونه تعارض منافع توسط نویسندگان وجود ندارد.

۹. تشکر و قدردانی

از مسئولین محترم مرکز اصلاح نژاد دام و بهبود تولیدات دامی کشور که داده‌های این پژوهش را در اختیار قرار دادند، تشکر و قدردانی می‌گردد.

۱۰. منابع

- Bergstra, J., & Yoshua, B. (2012). Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13, 281-305.
- Botchkarev, A. (2018). Performance metrics (error measures) in machine learning regression, forecasting and prognostics: properties and typology. *Interdiscipl. J. Inf. Knowl. Manag*, 14, 45-79. <https://arxiv.org/abs/1809.03006>.

- Carvalho, T., Soares, F., Vita, R., Francisco, R., Basto, J., & G. Soares Alcalá, S. (2019). A systematic literature review of machine learning methods applied to predictive maintenance. *Computers & Industrial Engineering*, 137, 106024.
- Ransom, C. J., Kitchen, N. R., Camberato, J. J., Carter, P. R., Ferguson, R. B., Fernández, F. G., ... & Shanahan, J. F. (2019). Statistical and machine learning methods evaluated for incorporating soil and weather into corn nitrogen recommendations. *Computers and electronics in agriculture*, 164, 104872. <https://doi.org/10.1016/j.compag.2019.104872>.
- Torshizi, M. E., Mashhadi, M. H., & Farhangfar, H. (2019). Different aspects of lactation persistency in dairy cows. *The Indian Journal of Animal Sciences*, 89(6), 607-614.
- Fadul-Pacheco, L., Delgado, H., & Cabrera, V. E. (2021). Exploring machine learning algorithms for early prediction of clinical mastitis. *International Dairy Journal*, 119, 105051.
- Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2), 83-85. <https://doi.org/10.1007/978-0-387-84858-7>.
- Horn, M., Steinwigger, A., Gasteiner, J., Podstatzky, L., Haiger, A., & Zollitsch, W. (2013). Suitability of different dairy cow types for an Alpine organic and low-input milk production system. *Livestock Science*, 153(1-3), 135-146.
- Hu, H., Mu, T., Ma, Y., Wang, X., & Ma, Y. (2021). Analysis of longevity traits in Holstein cattle: A review. *Frontiers in genetics*, 12, 695543.
- Imbayarwo-Chikosi, V. E., Ducrocq, V., Banga, C. B., Halimani, T. E., Van Wyk, J. B., Maiwashe, A., & Dzama, K. (2017). Estimation of genetic parameters for functional longevity in the South African Holstein cattle using a piecewise Weibull proportional hazards model. *Journal of Animal Breeding and Genetics*, 134(5), 364-372.
- Kuhn, M. (2019). The Caret Package. Available online: <http://topepo.github.io/caret/index.html>
- Lantz, B. (2015). *Machine learning with R* (Vol. 452). Birmingham: Packt publishing. ISBN 978-1-78216-214-8.
- Mammadova, N., & Keskin, I. (2013). Application of the support vector machine to predict subclinical mastitis in dairy cattle. *The Scientific World Journal*, 2013(1), 603897. <http://dx.doi.org/10.1155/2013/603897>.
- Pietersma, D., Lacroix, R., Lefebvre, D., & Wade, K. M. (2003). Performance analysis for machine-learning experiments using small data sets. *Computers and electronics in agriculture*, 38(1), 1-17.
- Robert, B., White, B. J., Renter, D. G., & Larson, R. L. (2009). Evaluation of three-dimensional accelerometers to monitor and classify behavior patterns in cattle. *Computers and electronics in agriculture*, 67(1-2), 80-84.
- Rostellato, R., Lora, I., Promp, J., Cassandro, M., Ducrocq, V., & Cozzi, G. (2022). Factors affecting true and functional productive lifespan in Italian Holstein-Friesian cows. *Italian Journal of Animal Science*, 21(1), 1268-1276.
- Schuster, J. C., Barkema, H. W., De Vries, A., Kelton, D. F., & Orsel, K. (2020). Invited review: Academic and applied approach to evaluating longevity in dairy cows. *Journal of Dairy Science*, 103(12), 11008-11024.
- Shahinfar, S., Khansefid, M., Haile-Mariam, M., & Pryce, J. E. (2021). Machine learning approaches for the prediction of lameness in dairy cows. *Animal*, 15(11), 100391.
- Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, 14(3), 199-222. <https://doi.org/10.1023/B:STCO.0000035301.49549.88>.
- Tamura, T., Okubo, Y., Deguchi, Y., Koshikawa, S., Takahashi, M., Chida, Y., & Okada, K. (2019). Dairy cattle behavior classifications based on decision tree learning using 3-axis neck-mounted accelerometers. *Animal Science Journal*, 90(4), 589-596.
- Pelt, M. V., Jong, G. D., & Veerkamp, R. F. (2017). Improving the genetic evaluation for longevity in the Netherlands. *Interbull bulletin* 51: 33-37, Tallinn, Estonia, August 25-28.
- Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., & Data, M. (2005, June). Practical machine learning tools and techniques. In *Data mining* (Vol. 2, No. 4, pp. 403-413). Amsterdam, The Netherlands: Elsevier.