DOI: 10.22092/VJ.2020.342148.1695
شماره ۱۳۲ پاییز ۱۴۰۰
تحقیقات دامپزشکی و فرآورده‌های بیولوژیک

تشخیص صفت فاصله زایش با استفاده از روشهای درخت تصمیم و ماشین بردار پشتیبان در گاوهای شیری هلشتاین

• **مریم منتظری نجف آبادی (نویسنده مسئول)**
دانشکده کشاورزی،دانشگاه یاسوج، ایران
• **محمدرضا بحرینی بهزادی**
گروه علوم دامی،دانشکده کشاورزی،دانشگاه یاسوج، ایران

تاریخ دریافت: ۱۳۹۸-۱۲-۱۰ تاریخ پذیرش: ۱۳۹۹-۰۲-۰۹
Email: montazeri_m12@yahoo.com

چکیده

کاهش خصوصیات تولیدمثلی در گاوهای شیری هلشتاین یک مشکل عمده در رابطه با پرورش گاوهای شیری است. رابطه منفی بین میزان تولید شیر و عملکرد تولیدمثلی در نژادهای مختلف وجود دارد. با تمرکز بر روی صفات تولیدمثلی مانند صفت فاصله زایش، تولید نیز در نسل‌های بعدی بهبود خواهد یافت. زیرا فاصله زایش یکی از فاکتورهایی است که میزان کارآمدی و کفایت تولیدمثل را نشان می‌دهد. استفاده از الگوریتم‌های هوشمند روش‌های یادگیری ماشین در بررسی سامانه‌های پیچیده رو به افزایش است و این روش‌ها نیز می‌توانند رهیافت مناسبی برای تحلیل داده‌های صنعت گاو شیری به‌حساب آیند. در پژوهش حاضر، بررسی امکان دسته‌بندی فاصله زایش در گاوهای شیری هلشتاین با استفاده از دو روش درخت تصمیم (الگوریتم‌های برجسته، جنگل تصادفی و بیز ساده) و ماشین بردار پشتیبان انجام شد. برای بررسی کارایی روش‌ها، از معیارهای صحت و جذر میانگین مربعات خطا استفاده شد. نتایج حاصل از این پژوهش نشان داد که روش‌های مختلف درخت تصمیم (برجسته) نسبت به روش ماشین بردار پشتیبان عملکرد بالاتری در امکان دسته‌بندی فاصله زایش داشت. سن مادر و تولید شیر، بیشترین ارتباط را با فاصله زایش داشتند. پژوهش حاضر اولین مطالعه با هدف تشخیص فاصله زایش با استفاده از روش‌های درخت تصمیم و ماشین بردار پشتیبان است که ممکن است اطلاعات در جهت درک بیشتر مدیریت فاصله زایش را بهبود دهد. عدم نیاز به برقراری هیچ پیش‌فرضی برای مدل‌سازی و تفسیر آسان نتایج مدل‌های درختی، دو مزیت اساسی آن است که به همین دلیل به‌نظر می‌رسد برای تحقیقات اصلاح نژاد دام مفید باشند.

کلمات کلیدی: درخت تصمیم‌گیری، دسته‌بندی، فاصله گوساله‌زایی، گاو شیری، ماشین بردار پشتیبان

Veterinary Researches & Biological Products No 132 pp: 99-106

Detection of Calving interval trait using decision tree and support vector machine methods in Holstein dairy cows

By:Montazeri Najafabadi, M., (Corresponding Author) Faculty of Agriculture, Yasouj University,Yasouj, Iran. and Bahreini Behzadi, M. R., Faculty of Agriculture, Yasouj University,Yasouj, Iran.

Received: 2020-02-29 **Accepted: 2020-04-28**

Email: montazeri_m12@yahoo.com

Decreased reproductive traits in cows are a major problem in raising dairy cows. There has been a negative relationship between milk production and reproductive performance in different breeds of dairy cows. By focusing on reproductive traits such as calving interval, production in later generations will be improved. Because calving interval is one of the factors that indicate the efficiency and adequacy of reproduction. Using intellectual algorithms of machine learning methods to investigate complex systems are growing and these algorithms could be assumed as right approach to analysis dairy cattle industry data. In the present work, was carried out to investigate the possibility of calving interval classification in Holstein dairy cattle using two methods of decision tree (algorithms of j48, random forest and Naive Bayes) and support vector machine. Were used accuracy and root mean square error to investigate the efficiency of methods. The results of this study showed that decision tree methods (j48) have a higher performance than support vector machine in classification of as calving interval. Calving age and milk production showed high amount of correlation with calving interval. The present work is the first study to detection of calving interval using decision tree and support vector machine methods that may provide information to greater understanding on calving interval management. Tree models don't require the establishment of no default for making model and feasibility of tree models results interpretation are two essential beneficiary of these models which for this reason seem to be useful for bovine breeding researches.

Keyword: Decision Tree, Classification, Calving interval, Dairy cow, Support Vector Machine

مقدمه

با توجه به همبستگی ژنتیکی نامطلوب بین صفات تولید شیر و باروری، توجه به صفات باروری در برنامه اصلاح‌نژادی توصیه شده است (۲۷). تولیدمثل از جمله سازه‌های کلیدی در تعیین بازده و سودآوری در صنعت پرورش گاوهای شیرده محسوب می‌شود، پس اگر بخواهیم سودمندی گاو شیری حداکثر شود لازم است بازده تولیدمثلی در سطح بالایی نگه داشته شود (۹،۲۴). بنابراین تولیدمثل را می‌توان به‌عنوان پایه و اساس استمرار بقای نسل در تمام موجودات زنده دانست که از برجسته‌ترین عوامل تعیین‌کننده کارآیی حیوانات مزرعه‌ای، بویژه گاوهای شیری به شمار می‌رود (۷). هدف اصلی در مدیریت تولیدمثل گاوهای شیرده، حداکثر نمودن تعداد زایمان‌ها در طول عمر گاوها و افزایش سود حاصل از فعالیت گاوداری می‌باشد (۴،۱۹،۲۴).

فاصله زایش یکی از صفات شاخص باروری در پرورش گاوهای شیری است و از نظر اقتصادی، میزان باروری در گاوهای شیرده معمولا با توجه به آن ارزیابی می‌شود. گزارش شده است که افزایش فاصله زایش باعث کاهش سودآوری می‌گردد (۲۷). با توجه به اینکه مهم‌ترین صفت ذکر شده برای توصیف بازده تولیدمثلی در هر گله، فاصله زایش است بنابراین این صفت یک سنجه مهم بازده خوب تولیدمثل در هر گله گاو شیری محسوب می‌شود. میانگین فاصله گوساله‌زایی ۴۰۰ روز برای گاوهای شیری در نظر گرفته شده است که برای کل شکم‌های زایش ۴۰۳±۶۵/۸ و دامنه آن از ۴۰۲/۹ روز برای گاوهای شکم اول تا ۴۱۲/۹ روز برای گاوهای شکم هفتم متغیر می‌باشد (۱۳،۲۶). در آخر دوره‌ی شیردهی، گاو زودتر از زمان مقرر و در حالی که تولید آن هنوز بالا است، خشک خواهد شد و احتمالا طول دوره‌ی خشکی، کمتر از آن خواهد بود که ترمیم پستان به‌طور کامل روی دهد. به همین دلیل تولید شیر در شیردهی بعدی کاهش خواهد یافت (۱۸،۲۳). فاصله گوساله‌زایی در گاوهای شیری، تولید شیر و عوامل تولید مانند هزینه‌های نگهداری مواد غذایی و مراقبت از حیوانات را تحت تاثیر قرار می‌دهد (۲،۵). باروری ضعیف باعث طولانی شدن فاصله زایش شده و اثرات منفی بیشتری را نسبت به باروری طبیعی از خود بر جای می‌گذارد (۱۱،۱۵). درخت تصمیم (Decision Tree:DT) درختی است که و دسته‌بندی استفاده می‌شود. این ساختار درخت مانند دارای گره‌ها، برگ‌ها و شاخه‌ها می‌باشد که جهت طبقه‌بندی داده‌ها به‌کار گرفته می‌شوند. الگوریتم جنگل تصادفی (Random Forest:RF) یک الگوریتم گروهی با مجموعه‌ای از درختان تصمیم است. در روش

درخت تصمیم‌گیری برجسته (J۴۸ یا C۴,۵) امکان هرس کردن آن معمولا با در نظر گرفتن یک حد آستانه پیاده‌سازی می‌شود (۲۲). در الگوریتم درخت بیز ساده (NBTree) مشخص‌شدن کلاس برگ‌های درخت بر پایه‌ی قانون بیز انجام می‌گیرد. درختان تصمیم از معیارهای مختلفی مثل بهره اطلاعات (Information Gain)، نسبت بهره (Gain Ratio)، شاخص جینی (Gini Index) و غیره برای انتخاب صفت خاصه استفاده می‌کند. ماشین بردار پشتیبان (Support Vector Machine) یک روش قدرت‌مند برای کلاسه‌بندی و رگرسیون است. روش ماشین بردار پشتیبان سعی دارد یک فراصفحه ایجاد نماید که فاصله هر کلاس را تا فراصفحه حداکثر کند. برای دسته‌بندی و تحلیل بیماری ورم پستان در گاوهای شیری از درخت تصمیم‌گیری استفاده شده است (۱۷). همچنین برای تشخیص و نظارت بر رفتارهای گاو شیری به‌منظور پژوهش‌های تغذیه‌ای، نشخوارکردن و خوابیدن آرام در گاوهای شیری، از روش‌های یادگیری ماشین (درخت تصمیم) استفاده شده است (۱۰،۲۰، ۲۵). گزارش‌هایی مربوط به استفاده از درخت تصمیم و ماشین بردار پشتیبان در تشخیص سخت‌زایی و بررسی سازه‌های موثر بر آن در گاو شیری نیز وجود دارد (۲۹،۲۸). از آنجایی که صفات تولیدمثلی بر اقتصاد گاوداری تاثیر دارند، لذا انجام پژوهش‌های علمی بر روی این صفات ضروری است. از آنجایی که تجزیه و تحلیل صفات تولیدمثلی مانند فاصله زایش به تصمیم‌گیری آگاهانه در قبال ضعف مدیریتی گله و مشکلات ناشی از آن کمک خواهد کرد، لذا هدف از انجام این تحقیق دسته‌بندی و پیش‌بینی صفت فاصله زایش در گاوهای شیری هلشتاین با استفاده از روش‌های درخت تصمیم و ماشین‌بردار پشتیبان می‌باشد تا مشخص شود کدام روش دارای دقت عملکرد بالاتری برای پیش‌بینی فاصله زایش بر اساس متغیرهای به‌کار رفته شده دارد. همچنین با اعمال روش‌های انتخاب ویژگی، مهم‌ترین متغیرهای تأثیرگذار بر فاصله زایش نیز استخراج شد.

مواد و روش‌ها

داده‌های مورد استفاده در این پژوهش مربوط به رکوردهای تولیدمثلی است که طی سال‌های ۱۳۸۸ تا ۱۳۹۶ توسط شرکت تعاونی وحدت گاوداران اصفهان جمع‌آوری شده بود. پایگاه داده شامل ۱۹۹۱۱۹ رکورد و ۵۰ ویژگی می‌باشد. صفات مورد مطالعه پس از استخراج از پایگاه داده دارای ۱۳ ویژگی (یک ویژگی هدف و ۱۲ ویژگی به عنوان ورودی) شامل مقدار شیر تولیدی در اوج شیردهی، تعداد شکم، جنس گوساله متولد شده، نوع تولد، سن مادر هنگام زایش، طول دوره آبستنی، فصل زایش، شیر تصحیح شده بر اساس چهار درصد چربی (FCM ٪۴)، نسبت چربی به پروتئین، وزن گوساله، روزهای باز و فاصله زایش به‌عنوان متغیر هدف در نظر گرفته شد. متغیر میانگین چربی شیر برای هر گاو با استفاده از نرم‌افزار SQL محاسبه شد و سپس مقدار شیر تصحیح شده بر اساس ۴ درصد چربی نیز با استفاده از رابطه (۱) به‌دست آمد:

(۱)

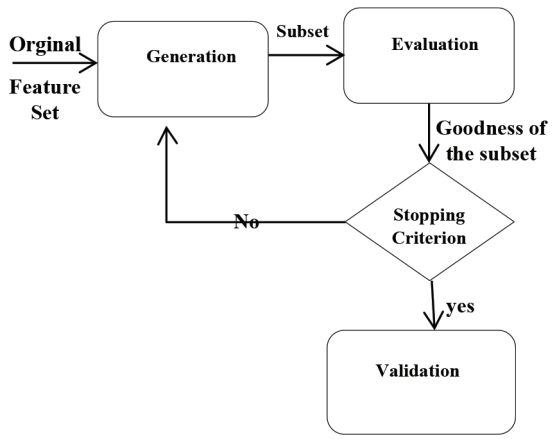
$$FCM4\% = 0.04\ (ML) + 15\ (FT)$$

در این رابطه ML مقدار شیر واقعی (خام) در هر دوره شیردهی (kg) و FT چربی شیر (kg) است.

تعدادی از رکوردها که دارای مقدار گمشده بودند حذف و ما بقی بوسیله میانگین جایگذاری شد. سپس داده‌ها برای استفاده نرمال‌سازی شدند. از سری داده اصلی با اندازه ۲۰۵۴۹ رکورد، سه نمونه تصادفی در اندازه‌های مختلف شامل ۱۰۲۲۰، ۳۰۰۰ و ۹۰۰ رکورد بوسیله‌ی نرم‌افزار SPSS ایجاد گردید و دسته‌بندی صفت فاصله زایش برای هر چهار سری داده انجام شد. به وضوح روشن شده است که انتخاب ویژگی نقش بسیار مهمی در شناسایی جایگاهای اتصال به عهده دارد. همواره دقت و صحت نتایج پیش‌بینی، ارتباط مستقیمی با تعداد و نوع ویژگی‌های انتخاب شده برای پیاده‌سازی الگوریتم‌های به‌کار برده شده دارد. انتخاب ویژگی فرآیندی است که زیر مجموعه‌ای از ویژگی‌های اصلی را انتخاب می‌کند. از طرفی از مهم‌ترین تکنیک‌های پیش‌پردازش داده به‌شمار می‌رود. بنابراین توسط این تکنیک ویژگی‌های اضافی و نامناسب شناسایی و حذف می‌شوند. داده با شش روش مهم الگوریتم انتخاب ویژگی (روش‌های Best-First، Greedy-Stepwise، Ranker، GeneticSearch، RandomSearch و RankSearch) جهت تعیین مهم‌ترین پارامترهای موثر بر فاصله‌ی زایش با استفاده از نرم‌افزار Weka واکاوی شد. در شکل ۱ فرآیند عمل الگوریتم انتخاب ویژگی نمایش داده شده است (۶). روش‌های مختلف انتخاب ویژگی در تلاش‌اند تا زیرمجموعه‌ای با کمینه تعداد ویژگی ممکن برای پیش‌بینی هدف موردنظر پیدا کنند. در تمام این روش‌ها بر اساس کاربرد و نوع تعریف، زیر مجموعه‌ای به‌عنوان جواب انتخاب می‌شود که بتواند مقدار یک تابع ارزیابی را بهینه کند. روند انتخاب ویژگی در تمامی روش‌ها به بخش‌های متوالی تقسیم می‌شود:

تابع تولیدکننده: این تابع زیرمجموعه‌های کاندید را برای روش موردنظر پیدا می‌کند.

تابع ارزیابی: زیرمجموعه‌ی موردنظر را بر اساس روش داده شده، ارزیابی می‌کند و یک عدد را به‌عنوان میزان خوب بودن زیرمجموعه‌ی موردنظر باز می‌گرداند. این مقدار با بهترین زیرمجموعه‌ی قبلی مقایسه می‌شود و اگر زیرمجموعه‌ی جدید بهتر از زیرمجموعه‌های قدیمی باشد، زیرمجموعه‌ی جدید به‌عنوان زیرمجموعه‌ی بهینه، جایگزین قبلی



شکل ۱- فرآیند انتخاب ویژگی.

می‌شود. روش‌های مختلف تلاش می‌کنند تا زیرمجموعه‌ای را پیدا کنند که این مقدار را بهینه کند. بنابراین تابع تولیدکننده که نقش جستجوگر دارد، زیرمجموعه‌های مختلف را به ترتیب تولید می‌کند تا بوسیله‌ی تابع ارزیابی، مورد ارزیابی قرار گیرد.

شرط خاتمه: برای تصمیم‌گیری در مورد زمان توقف الگوریتم.

تابع تعیین اعتبار: به کمک تابع تعیین اعتبار مشخص می‌شود که آیا زیرمجموعه‌ی انتخاب شده معتبر است یا خیر.

تابع تولیدکننده می‌تواند از یکی از حالت‌های زیر شروع به‌کار کند:

بدون ویژگی: در این حالت ویژگی‌ها به‌ترتیب به مجموعه اضافه می‌شوند و زیرمجموعه‌های جدید را تولید می‌کند. این عمل آنقدر تکرار می‌شود تا زیرمجموعه‌ی مورد نظر حاصل شود.

با مجموعه تمام ویژگی‌ها: از یک مجموعه شامل تمام ویژگی‌ها شروع می‌شود و به مرور و در طی اجرای الگوریتم، ویژگی‌ها حذف می‌شود تا به زیرمجموعه‌ی دلخواه برسد.

با یک زیر مجموعه تصادفی: این حالت همانند حالت قبلی است ولی از یک زیر مجموعه تصادفی شروع می‌شود.

به‌طور کلی روش‌های مختلف، انتخاب ویژگی را بر اساس نوع جستجو به دسته‌های مختلفی تقسیم‌بندی می‌کنند که در بعضی از روش‌ها، تمام فضای ممکن جستجو می‌گردد و در سایر روش‌ها که می‌تواند مکاشفه‌ای و یا جستجوی تصادفی باشد، در ازای از دست دادن مقداری از کارایی، فضای جستجو کوچک‌تر شود.

روش‌های انتخاب ویژگی

روش CfsSubsetEval

در این روش، ارزش پارامترها با استفاده از همبستگی ارزیابی می‌شود. بهترین زیرمجموعه‌ی ویژگی، دارای ویژگی‌هایی با بیشترین همبستگی برای پیش‌بینی ویژگی هدف است. اما درجه همبستگی بین خود ویژگی‌های موجود در زیر مجموعه کم می‌باشد. رابطه (۲) برای به‌دست آوردن درجه شایستگی این زیرمجموعه‌ها استفاده می‌شود (۱۲):

(۲)

$$\text{شایستگی} = \frac{\text{درجه‌ی همبستگی ویژگی مورد نظر با ویژگی هدف}}{\text{درجه‌ی همبستگی ویژگی مورد نظر با ویژگی‌های موجود در زیر مجموعه}}$$

روش ChiSquaredAttributeEval

این الگوریتم برای رتبه‌بندی ویژگی‌ها از معیار کای اسکوئر (Chi-Squared) استفاده می‌کند. ویژگی‌هایی که رتبه‌ی بالاتری دارند، از قدرت جداکنندگی بیشتری نیز برخوردار هستند.

برای انتخاب بهترین روش از میان سایر روش‌های انتخاب ویژگی به‌منظور موثرترین متغیرها بر صفت فاصله زایش از دو معیار میانگین کل کلاس‌های نادرست پیش‌بینی شده توسط مدل (OAMS) و میانگین کل نسبت موارد درست که مدل آنها را به درستی به عنوان کلاس درست پیش‌بینی کرده است (OATPrate) برای تعیین میزان خطای هر یک از روش‌ها از رابطه‌های ۳ و ۴ استفاده گردید (۱۶):

(۳)

$$OAMS = \frac{\sum_{i=1}^{n} AMS}{N}$$

(۴)

$$OATPrate = \frac{\sum_{i=1}^{n} ATPrate}{N}$$

در این رابطه‌ها، دو معیار AMS و ATPrate به ترتیب بیانگر میانگین تعداد کلاس‌های نادرست پیش‌بینی شده توسط مدل، و میانگین نسبت موارد درستی توسط پیش‌بینی درست مدل و N تعداد کل داده‌های مورد استفاده در آزمایش را نشان می‌دهد.

تابع کرنل (Kernel) تابعی است که برای ایجاد ماشین‌هایی با انواع مختلفی از سطوح تصمیم‌گیری غیرخطی در فضای داده‌ها، ضرب‌های داخلی تولید می‌کند که مهم‌ترین پارامتر آموزشی این روش، نوع کرنل است که از رابطه (۵) محاسبه می‌گردد.

(۵)

$$K(X, Xi) = (X \times Xi + 1)d$$

در رابطه بالا، تابع K (X , Xi)تابع کرنل نامیده می‌شود. d درجه کرنل چند جمله‌ای و علامت × نشان‌دهنده‌ی ضرب داخلی است. در روش ماشین بردار پشتیبان، تابع کرنل از نوع چند جمله‌ای (poly nominal) و برای تعمیم این روش به حالت چند کلاسه از روش یکی در مقابل همه استفاده شده است. یکی از روش‌های مورد استفاده برای تعمیم ماشین‌های بردار پشتیبان در حالت چند کلاسه، روش یکی در مقابل همه می‌باشد که در این روش هر کلاس در مقابل سایر کلاس‌ها آموزش داده می‌شود سپس برای داده‌های مربوط به خود، کلاس برچسب +۱ و داده‌های سایر کلاس‌ها، بر چسب -۱ لحاظ می‌شود. پس از آموزش کلاس‌ها، در مرحله‌ی تست، هر یک از نمونه‌های تست به تمامی N ماشین بردار پشتیبان اعمال می‌شود و کلاس برنده کلاسی است که ماشین بردار پشتیبان آن بیشترین میزان خروجی را داشته باشد. همچنین از دو معیار نسبت بهره و بهره اطلاعات به منظور انتخاب نقطه شکست شاخه در مدل‌سازی به روش درخت تصمیم استفاده شد. در روش معیار بهره اطلاعات برای هر صفت این معیار بر اساس رابطه (۶) محاسبه شده و هر صفتی که بیشترین مقدار را کسب کرد به عنوان ریشه درخت انتخاب می‌شود.

(۶)

$$IG(S,A) = Entropy(S) - \sum |S_V| / |S| Entropy(S_V)$$

در رابطه (۶)، Sv زیر مجموعه‌ای از S (مجموعه داده مورد نظر)، است. در حالتی که متغیر هدف دارای چند کلاس باشد رابطه آنتروپی از رابطه (۷) محاسبه می‌شود:

(۷)

$$Entropy(S) = - \sum P_i \times \log(p_i)$$

در رابطه (۷)، Pi معرف احتمال اینکه نمونه‌ای از داده‌ها متعلق به کلاس iام باشد.

پس از ویرایش، داده‌ها با استفاده از نرم‌افزار WEKA واکاوی و مجموعه داده به دو بخش آموزش (۷۰ درصد) و آزمون (۳۰درصد) تقسیم شد. برای مقایسه مدل‌ها از جذر میانگین مربعات خطا (RMSE) و معیار

صحت (Accuarcy) استفاده گردید :

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (۸)$$

$$Accuarcy = \frac{TP + TN}{TP+FP+FN+TN} \quad (۹)$$

در رابطه (۸) y_i مقدار مشاهده شده رکورد فاصله زایش، y^_i مقدار پیش‌بینی شده آن توسط مدل و n تعداد رکوردها می‌باشند. پارامترهای موجود در معیار صحت رابطه (۸) با استفاده از ماتریس اغتشاش که در جدول ۱ ارائه شده است محاسبه می‌شود:

سطرهای این ماتریس به مقادیر واقعی و ستون‌های آن به مقادیر پیش‌بینی شده اشاره دارند. در حالت ساده وقتی فرض شود که مشخصه‌ای را که قرار است مقدار آن پیش‌بینی شود و داده‌ها بر اساس آن پیش‌بینی می‌شوند، فقط شامل یکی از دو کلاس مثبت و منفی می‌باشد که به ترتیب، به اختصار با P و N نام برده می‌شود. در این حالت بسته به اینکه مقدار این مشخصه واقعا کدامیک از این دو حالت بوده است و مقدار پیش‌بینی مدل، درست یا نادرست (به اختصار T یا F) باشد، یک ماتریس دو در دو به صورت جدول ۱ ایجاد می‌شود. برای تحلیل داده‌های چند کلاسه (شامل بیش از دو کلاس) سطر و ستون‌های ماتریس اغتشاش بسط داده می‌شود.

جدول ۱- ماتریس اغتشاش برای نتایج مختلف پیش‌بینی دو دسته.

		کلاس پیش‌بینی شده توسط روش دسته‌بندی	
		دسته ۱	دسته ۲
کلاس واقعی	کلاس ۱	TP (True Positive)	FN (False Negative)
	کلاس ۲	FP (False Positive)	TN (True Negative)

TP: مثبت صحیح، FP: مثبت کاذب، FN: منفی کاذب، TN: منفی صحیح.

جدول ۲- موثرترین عوامل تاثیرگذار بر فاصله زایش به‌دست آمده با روش‌های مختلف انتخاب ویژگی.

روش	مدل	اندازه نمونه	فراسنجه
Greedy-Stepwise	CFS-Subset-Eval	۹۰۰	سن مادر هنگام زایش، شیر تصحیح شده بر اساس ۴٪ چربی، نسبت چربی به پروتئین، وزن گوساله
Best-First	CFS-Subset-Eval	۳۰۰۰	سن مادر هنگام زایش، روزهای باز
Best-First	ConsistencySubsetEval	۱۰۲۲۰	سخت‌زایی، کد گله، سن مادر هنگام زایش، تعداد شکم، طول دوره آبستنی، ماه زایش، نوع تولد، شیر تصحیح شده بر اساس ۴٪ چربی، نسبت چربی به پروتئین، جنسیت گوساله، وزن گوساله، روزهای باز
Ranker	ChiSquaredAttributeEval	۲۰۵۴۹	سن مادر هنگام زایش، کد گله، شیر تصحیح شده بر اساس ۴٪ چربی، روزهای باز، سال زایش، سخت‌زایی، طول دوره آبستنی، تعداد شکم، فصل زایش، نوع تولد، جنسیت گوساله، وزن گوساله، نسبت چربی به پروتئین، مقدار شیر تولیدی در اوج شیردهی

نتایج و بحث

موثرترین عوامل تاثیرگذار نهایی بر فاصله زایش که با روش‌های مختلف الگوریتم انتخاب ویژگی به‌دست آمده به تفکیک روش و مدل برای هر نمونه در جدول ۲ آورده شده است.

جدول ۳ نتایج الگوریتم‌های مختلف را روی مجموعه داده‌ها بر اساس بیشترین مقدار صحت (Accuarcy) و کمترین مقدار جذر میانگین مربعات خطا (RMSE) نشان می‌دهد.

الگوریتم درخت تصمیم‌گیری برجسته (J۴۸)، در دقت پیش‌بینی فاصله زایش، بالاتر از سایر الگوریتم‌های مقایسه قرار می‌گیرد و با ۶۹/۳۳ درصد، بهترین و بیشترین مقدار Accuarcy و کمترین مقدار RMSE را روی نمونه ۹۰۰ رکوردی نشان می‌دهد. همچنین برای نمونه‌های دیگر نیز این الگوریتم نسبت به سایر الگوریتم‌های دیگر دقت پیش‌بینی بالاتری داشته است. روش SVM با کمترین مقدار (۴۰/۱ درصد) بدترین نتیجه را روی نمونه ۳۰۰۰ رکوردی ارائه می‌دهد و پس از آن الگوریتم جنگل تصادفی (RF) روی داده اصلی (۲۰۵۴۹) و نمونه‌های گرفته شده از آن (۳۰۰۰ و ۱۰۲۲۰ رکوردی) دقت پیش‌بینی کمتر و عملکرد پایینی دارد. از طرفی الگوریتم ماشین بردار پشتیبان (SVM)، تنها روی نمونه ۳۰۰۰ رکوردی عملکرد پایینی دارد ولی عملکرد این الگوریتم روی داده ۲۰۵۴۹ و نمونه ۱۰۲۲۰ در مقایسه با الگوریتم جنگل تصادفی (RF) بسیار بهتر و بیشتر است. بر اساس نتایج این پژوهش، می‌توان چنین استنباط نمود که عملکرد پایین الگوریتم ماشین بردار پشتیبان (SVM) به دلیل تعداد متغیر های مورد استفاده در این الگوریتم است. زیرا تعداد متغیرهای مورد استفاده در این الگوریتم روی نمونه ۳۰۰۰ رکوردی کمتر از نمونه‌های دیگر است.

از طرفی تعداد متغیرهای کمتر، تاثیر بیشتری بر دقت پیش‌بینی روش درخت تصمیم‌گیری در مقایسه با روش ماشین بردار پشتیبان داشته است. در مدل‌سازی صفت سخت‌زایی با استفاده از روش‌های درخت تصمیم و ماشین بردار پشتیبان، درخت تصمیم عملکرد بالایی را برای دسته‌بندی صفت سخت‌زایی داشت و سن مادر هنگام زایش و پیک تولید شیر از متغیرهای تاثیر گذار بر این صفت بود (۳). نتایج حاصل از پژوهش حاضر نشان داد که روش درخت تصمیم با داشتن دقتی نسبتا بالا، کارایی قابل قبولی در مدل‌سازی صفت فاصله گوساله‌زایی دارد که با توجه به اهمیت صفات تولیدمثلی به نظر می‌رسد که از نتایج این پژوهش بتوان به مدیریت بهتر صفات تولیدمثلی در گله‌های گاو شیری استفاده کرد. اما در مطالعه‌ای که توسط آدمزاک و همکاران (۱) صورت گرفت، نشان داده شد که استفاده از شبکه عصبی مصنوعی در مقایسه با درخت تصمیم به منظور پیش‌بینی دلایل مختلف حذف بر اساس داده‌های عملکرد طول عمر مناسب‌تر می‌باشد. در پژوهش‌های صورت گرفته توسط دولیالا و همکاران (۸) رفتار فحلی در گاوهای شیری با استفاده از ماشین بردار پشتیبان دسته‌بندی شد. از الگوریتم‌های درخت تصمیم برجسته، جنگل تصادفی با معیار بهره اطلاعات و شبکه‌های بیزی برای پیش‌بینی نتایج تلقیح در گاوهای شیری هلشتاین، استفاده شد که بهترین الگوریتم، جنگل تصادفی بود و نشان داد، متغیرهای میانگین نرخ تلقیح در گله، شیوع کتوز، روزهای باز، دوره شیردهی و ورم پستان ارتباط بیشتری با آبستنی دارد (۲۵). در مطالعه‌ای که روی موفقیت تخمک‌گذاری گاو شیری در یک دوره شیردهی صورت گرفت، درختان تصمیم عملکرد تقریبا خوبی را نسبت به ماشین بردار پشتیبان از خود نشان داد. در این پژوهش از

جدول ۳- نتایج الگوریتم‌های مختلف بر اساس معیار صحت (درصد) و RMSE روی مجموعه آزمون.

الگوریتم‌ها	اندازه نمونه				
	معیارها	۲۰۵۴۹	۱۰۲۲۰	۳۰۰۰	۹۰۰
J48	Accuarcy	۵۵/۳۶	۵۶/۵۳	۵۲/۹۳	۶۹/۳۳
	RMSE	۰/۳۸	۰/۳۷	۰/۳۹۲	۰/۳۲
NB	Accuarcy	۵۵/۰۲	۵۶/۲۷	۴۸/۷۶	۶۰/۴۴
	RMSE	۰/۳۷۴	۰/۳۷۲	۰/۴۲۳	۰/۳۶
RF	Accuarcy	۴۷/۳۳	۴۸/۲۲	۴۴/۱۶	۶۸/۲۲
	RMSE	۰/۴۱۶	۰/۳۹۷	۰/۴۱۲	۰/۳۳
SVM	Accuarcy	۵۳/۲۱	۵۲/۵۵	۴۰/۱	۵۱/۵۵
	RMSE	۰/۴۰۱	۰/۴۰۸	۰/۴۳۹	۰/۴۱

Accuarcy: صحت، RMSE : جذر میانگین مربعات خطا.

J۴۸: درخت تصمیم‌گیری برجسته، NB: درخت بیز ساده، RF: جنگل تصادفی، SVM: ماشین بردار پشتیبان.

تابع کرنل چندجمله‌ای و پایه شعاعی برای آموزش ماشین بردار پشتیبان و از معیار بهره اطلاعات برای آموزش درخت تصمیم استفاده شد (۱۴). برای برآورد پارامترهای شیر تولیدی در یک مدل اقتصادی مربوط به یک گاوداری در اروپا از ماشین بردار پشتیبان استفاده شد که نشان داد از این روش می‌توان برای برآورد پارامترهای مدل خطی و غیرخطی استفاده کرد (۲۱). در مطالعه‌ای که بر روی میزان سخت‌زایی در تلیسه‌ها صورت گرفت، تعداد هشت متغیر با ماهیت‌های متفاوت تولیدی، هورمونی و مدیریتی به‌عنوان متغیرهای مورد استفاده در روش درخت تصمیم و ماشین بردار پشتیبان به‌کار برده شدند. هدف این بود که مشخص گردد چه تعداد از متغیرها باعث بروز سخت‌زایی می‌گردند. نشان داده شد که متغیرهای طول دوره آبستنی، شاخص وضعیت بدن و سن زایش بیشترین تاثیر و ارتباط را با بروز سخت‌زایی بوسیله‌ی درخت تصمیم داشتند. در این مطالعه سه معیار شاخص جینی (Gini Index)، کای اسکوئر (chi-square) و جی اسکوئر (G-square) به‌عنوان انشعاب شاخه‌ها و کرنل خطی به ترتیب برای درخت تصمیم و ماشین بردار پشتیبان به‌کار برده شد (۲۹). همچنین میزان سخت‌زایی در گاوهای شیری با استفاده از درختان طبقه‌بندی ارتقاءدهنده پیش‌بینی شد و فاکتورهای کلیدی مرتبط با سخت‌زایی طول دوره آبستنی، شاخص وضعیت بدن، سن زایش، ژن هلشتاین-فریزن، روزهای باز و شیر تصحیح شده بر اساس ۴ درصد چربی است (۲۸).

نتیجه‌گیری

به‌طور کلی نتایج این پژوهش نشان داد که الگوریتم درخت تصمیم‌گیری برجسته (J48) از میان روش‌های درخت تصمیم و دیگر روش‌های مورد استفاده در این پژوهش، توانسته است پیش‌بینی خوبی از فاصله زایش ارائه دهد. این روش عملکرد بالایی برای دسته‌بندی فاصله زایش بر اساس متغیرهای به‌کار رفته در این مطالعه دارد. همچنین سن مادر هنگام زایش، شیر تصحیح شده بر اساس چهار درصد چربی، تعداد شکم، وزن گوساله، نسبت چربی به پروتئین، روزهای باز و جنسیت گوساله، مهم‌ترین متغیرهای تاثیرگذار بر فاصله زایش می‌باشند. در این مطالعه مشخص شد که افزایش تعداد متغیر بر روی عملکرد ماشین بردار پشتیبان (SVM) تاثیرگذار است و عملکرد آن تغییر می‌کند. پیشنهاد می‌شود از سایر روش‌های انتخاب ویژگی مانند روش‌هایی که تابع ارزیابی مبتنی بر خطای طبقه‌بندی کننده و توابع مختلف تولید کننده است استفاده شود.

تشکر و قدردانی

داده‌های مورد استفاده در این تحقیق توسط شرکت تعاونی وحدت گاوداران اصفهان، تهیه شده است که بدینوسیله از مسئولین شرکت مزبور تشکر و قدردانی می‌شود.

منابع مورد استفاده

1.Adamczyk, K., D. Zaborski, W. Grzesiak, J. Makulska and W. Jagusiak. 2016. Recognition of culling reasons in Polish dairy cows using data mining methods. *Computers and Electronics in Agriculture*, 127: 26-37.

2. Arbel, R., Y. Bigun, E. Ezra, H. Sturman and D. Hojman. 2001. The effect of extended calving intervals in high lactating cows on milk production and profitability. *Journal of Dairy Science*, 84: 600–608.

3.Bahreini Behzadi, M., R. M. Montazeri NajafAbadi. 2018. Detection of dystocia trait using support vector machine and decision tree methods in Holstein dairy cows. In: Proceeding of 8th International Animal Science. Iran, Kordestan. 4-6. (In Farsi).

4.Berry, D., B. Harris, A.Winkelman and W. Montgomerie. 2005. Phenotypic associations between traits other than production and longevity in New Zealand dairy cattle. *Journal of Dairy Science*, 8: 2962-2974.

5.Dalcq, A., C. Y. Beckers P. Mayeres, E. Reding, B. Wyzen, F. Colinet, ... and H. Soyeurt. 2018. The feeding system impacts relationships between calving interval and economic results of dairy farms. *Journal animal Science*, 12(8): 1662-1671.

6.Dash, M. and H. Liu. 2003. Consistency-based search in feature selection. *Artificial Intelligence*, 151: 155–176.

7.Dematawewa C. and P. Berger. 1998. Genetic and Phenotypic Parameters for 305-Day Yield, Fertility, and Survival in Holsteins1. *Journal of Dairy Science*, 10: 2700-2709.

8.Dulyala, R., S. Kuankid, T. Rattanawong and A. Aurasopon. 2014. Classification system for estrus behavior of cow using an accelerometer. In Signal and Information Processing Association Annual Summit and Conference (APSIPA), Asia-Pacific, 14-15.

9.Ghavi Hossein- Zadeh, N. 2014. Effect of dystocia on the productive performance and calf stillbirth in Iranian Holsteins. *Journal of Agricultural Science and Technology*, 1: 69-78.

10.Gonzalez, L. A., G. J. Bishop, R. N. Hurley Handcock and C. Crossman. 2015. Behavioral classification of data from collars containing motion sensors in grazing cattle. *Computers and Electronics in Agriculture,* 110: 91-102.

11.Groenendaal, H., D.T. Galligan and H. A. Mulder. 2004. An economic spreadsheet model to determine optimal breeding and replacement decisions for dairy cattle. *Journal of Dairy Science*, 87: 2146–2157.

12.Hall, M. A. 1998. Correlation-based Feature Subset Selection for Machine Learning. Thesis submitted in partial fulfillment of the requirements of the degree of Doctor of Philosophy at the University of Waikato Hamilton, New Zealand.

13.Hare, E. H. D. N., H. D. Norman and J. R. Wright. 2006. Trends in calving ages and calving intervals for dairy cattle breeds in the United States. *Journal of dairy science*, 89(1): 365-370.

14.Hempstalk, K., S. McParland and D. P. Berry. 2015. Machine learning algorithms for the prediction of conception success to a

given insemination in lactating dairy cows. *Journal of dairy science*, 98(8): 5262-5273.

15.Inchaisri, C., R. Jorritsma, P. L. A. M. Vos, G.C. van der Weijden and H. Hogeveen. 2011. Analysis of the economically optimal voluntary waiting period for first insemination. *Journal of dairy science*, 74: 835–846.

16.Kargar, S. and M. Mokarram. 2016. Use of feature selection algorithm to determine the most important factors affecting milk fat percentage of Holstein cows. *Iranian Journal of Ruminant Research*, 4(4): 154-166. (In Farsi).

17.Kim, T and C. W. Heald. 1999. Inducing inference rules for the classification of bovine mastitis. *Computers and Electronics in Agriculture*, 23(1): 27-42.

18.Lehmann, J. O., J. G. Fadel, L. Mogensen, T. Kristensen, C. Gaillard and E. Kebreab. 2016. Effect of calving interval and parity on milk yield per feeding day in Danish commercial dairy herds. *Journal of dairy science*, 99(1): 621-633.

19.Maizon, D., P. Oltenacu, Y. Grohn, R. Strawderman and U. Emanuelson. 2004. Effects of diseases on reproductive performance in Swedish Red and White dairy cattle. *Preventive veterinary medicine*, 1: 113-126.

20.Martiskainen و P., M. Jarvinen, J. P. Skon, J. Tiirikainen, M. Kolehmainen and J. Mononen. 2009. Cow behaviour pattern recognition using a three-dimensional accelerometer and support vector machines. *Applied animal behaviour science*, 119(1-2): 32-38.

21.Poldaru, R., J. Roots and A. H. Viira. 2005. The estimation of the econometric model of milk yield per cow: a support vector machine regression approach.Congress on it in Agriculture Operations Research, 166-167.

22.Qin, Z and J. Lawry. 2005. Decision tree learning with fuzzy labels. *Information Sciences*, 172: 91 129.

23.Royal, M. D., J. E. Pryce, J. A. Woolliams and A. P. F. Flint. 2002. The genetic relationship between commencement of luteal activity and calving interval, body condition score, production, and linear type traits in Holstein-Friesian dairy cattle. *Journal of dairy science*, 85(11): 3071-3080.

24.Sewalem, A., F. Miglior, G. Kistemaker, P. Sullivan and B. Van Doormaal. 2008. Relationship between reproduction traits and functional longevity in Canadian dairy cattle. *Journal of Dairy Science*, 91:1660–1668.

25.Shahriar, M. S., D. Smith, A. Rahman, M. Freeman, J. Hills, R. Rawnsley, ... and G. Bishop Hurley. 2016. Detecting heat events in dairy cows using accelerometers and unsupervised learning. *Computers and Electronics in Agriculture* 128: 20-26.

26.Silva, H. M., C. J. Wilcox, W. W. Thatcher, R. B. Becker and D. Morse. 1992. Factors Affecting Days Open, Gestation Length and Calving Interval in Florida Dairy Cattle1. *Journal of Dairy Science*, 75(1): 288-293.

27.Toghiani, S., A. A. Shadparvar, M. Moradi Shahrbabak and M. Dadpasand Taromsari. 2009. Genetic analysis of reproduction traits and their relationship with conformation traits in Holstein cows. *Livestock Production Science*, 125: 84-87.

28.Zaborski, D., W .Grzesiak, K. Kotarska, I. Szatkowska and M. Jedrzejczak. 2014. Detection of difficult calvings in dairy cows using boosted classification trees. *Indian Journal of Animal Research*, 48(5): 452-458.

29.Zaborski, D., W. Grzesiak and R. Pilarczyk. 2016. Detection of difficult calvings in the Polish Holstein-Friesian Black-and-White heifers. *Journal of applied animal research*, 44(1): 42-53.