University of Tehran Press

Journal of Animal Production

Online ISSN: 2382-994X

Homepage: https://jap.ut.ac.ir/

Using machine learning algorithms to predict the occurrence of clinical mastitis in Holstein cows

Mohammad-Taghi Fayazi-Kia[1] | Mohammad Dadpasand[2]✉ | Hamideh Keshavarzi[3]

1. Department of Animal Science, School of Agriculture, Shiraz University, Shiraz, Iran. E-mail: fayazi.mt@gmail.com
2. Corresponding Author, Department of Animal Science, School of Agriculture, Shiraz University, Shiraz, Iran. E-mail: dadpasand@shirazu.ac.ir
3. Agriculture and Food, Commonwealth Scientific and Industrial Organization (CSIRO), Armidale, NSW 2350, Australia. E-mail: Hamideh.Keshavarzi@csiro.au

Article Info

Article type:
Research Article

Article history:
Received: 2 October 2022
Received in revised form:
3 June 2023
Accepted: 6 June 2023
Published online: 6 July 2023

Keywords:
Dairy cow,
Machine learning,
Mastitis,
Prediction,
Sampling.

ABSTRACT

Introduction: Mastitis is one of the most frequent and costly diseases of the dairy cattle industry and causes many economic losses, which negatively affects milk yield and composition, fertility, longevity and welfare of cows. The best solution for reducing the economic and biological consequences is early and accurate prediction of mastitis based on indicator factors. So far, various statistical methods have been used to predict mastitis such as linear and multiple regression, and threshold models. Machine learning is another method that has recently widely been used to predict farm profitability, reproductive traits, longevity and abortion in dairy cow. Machine learning is defined as a set of methods for automatically finding patterns in data and then using those patterns to predict possible future data.

Material and Methods: In this research, the performance of four machine learning algorithms including random forest, decision tree, Naïve Bayes and logistic regression and two sampling methods, over-sampling and under-sampling, were compared to predict risk of clinical mastitis based on data collected in two Holstein dairy herds in Isfahan province. Final dataset included 393504 records on cows calved during 2007 to 2017 of which 13653 cases (3.47%) were infected and 379851 cases (96.53%) were healthy. Factors related to mastitis, including parity, daily milk production, calving season, lactation stage, history of mastitis and somatic cell score. After editing the data with SQL Server software, the modeling process was implemented to predict mastitis using WEKA 3.8 software. The performance of algorithms (accuracy, sensitivity, specificity, and AUC) in predicting infected cases and distinguishing them from healthy cases was evaluated according to the preprocessing method used. The sampling techniques used included Under Sampling (SpreadSubSampling) and Synthetic minority oversampling technique (SMOTE).

Results and Discussion: Results showed that the best performance among the algorithms was related to the random forest in the case of using the low-sampling method with the accuracy, sensitivity, detection and AUC rates of 84.30%, 94.80%, 73.80% and 90.90%, respectively. In the case of not using sampling, the power to detect sick cases (sensitivity in percentage) in random forest, decision tree, Naïve Bayes and logistic regression algorithms was 1.67, 0, 12.29 and 2.06, respectively, which compared to sampling was considerably weaker. This was due to the unbalanced number of cases in two classes, healthy and sick, and indicated the necessity of using sampling methods. The decision tree algorithm in the case of low-sampling method with a small difference after the random forest has the best performance with accuracy, sensitivity, detection and AUC 84.00%, 94.20%, 73.90% and 90%, respectively. Comparing the models obtained from the four algorithms Decision Tree, Logistic, Naïve Bayes and Random Forest in three modes without preprocessing, with SpreadSubSample preprocessing and with SMOTE preprocessing, among the preprocessing modes, preprocessing by SMOTE method can significantly improve the performance of the algorithms. Among the algorithms that were pre-processed with this method, the Random Forest algorithm has shown the best performance with an accuracy of 99.2% and an area under ROC curve (AUC) of 0.99. Decision Tree algorithm has performed very close to Random Forest with accuracy of 98.9 and AUC of 0.99. Likewise, Naïve Bayes algorithm with accuracy and AUC of 0.92 and 82.9 and Logistic algorithm with accuracy and AUC of 83.7 and 0.91, respectively had acceptable performances after the other two algorithms.

Conclusion: Due to the high performance of the Random Forest algorithm using the SMOTE preprocessing method, in predicting mastitis cases, the use of this model can be suggested to predict cases of mastitis in dairy cattle herds, especially in herds with high rates of mastitis. Because of higher computational cost of random forest compared to random tree, in large dataset, decision tree probably should be a better choice.

Cite this article: Fayazi-Kia, M. T., Dadpasand, M., & Keshavarzi, H. (2023). Using machine learning algorithms to predict the occurrence of clinical mastitis in Holstein cows. *Journal of Animal Production*, 25 (2), 123-132.
DOI: https://doi.org/10.22059/jap.2023.349388.623708

Publisher: University of Tehran Press.
DOI: https://doi.org/10.22059/jap.2023.349388.623708

انتشارات دانشگاه تهران

تولیدات دامی

شاپا الکترونیکی: ۲۳۸۲-۹۹۴x

Homepage: https://jap.ut.ac.ir/

استفاده از الگوریتم‌های یادگیری ماشین برای پیش‌بینی وقوع ورم پستان بالینی در گاوهای هلشتاین

محمدتقی فیاضی‌کیا[۱] | محمد دادپسند[۲]✉ | حمیده کشاورزی[۳]

۱. بخش علوم دامی، دانشکده کشاورزی، دانشگاه شیراز، شیراز، ایران. رایانامه: fayazi.mt@gmailcom

۲. نویسنده مسئول، بخش علوم دامی، دانشکده کشاورزی، دانشگاه شیراز، شیراز، ایران. رایانامه: dadpasand@shirazu.ac.ir

۳. مرکز پژوهش‌های علمی و صنعتی غذا و کشاورزی مشترک المنافع (CSIRO)، آرمیدل، نیوساوت ولز، استرالیا. رایانامه: Hamideh.Keshavarzi@csiro.au

اطلاعات مقاله

نوع مقاله: مقالهٔ پژوهشی

تاریخ دریافت: ۱۴۰۱/۰۷/۱۰

تاریخ بازنگری: ۱۴۰۲/۰۳/۱۳

تاریخ پذیرش: ۱۴۰۲/۰۳/۱۶

تاریخ انتشار: ۱۴۰۲/۰۴/۱۵

کلیدواژه‌ها:
پیش‌بینی،
گاو شیری،
نمونه‌گیری،
ورم پستان،
یادگیری ماشین.

چکیده

در این پژوهش، از چهار الگوریتم جنگل تصادفی، درخت تصمیم، بیز ساده و رگرسیون لجستیک برای پیش‌بینی بیماری ورم پستان براساس داده‌های دو گله گاو شیری هلشتاین استفاده شد. به‌دلیل نامتوازن بودن تعداد موارد بیمار و سالم از دو روش بیش‌نمونه‌برداری و کم‌نمونه‌برداری استفاده شد. متغیرهای مرتبط با ورم پستان، شامل نوبت زایش، تولید شیر روزانه، فصل زایش، مرحله شیردهی، سابقه ورم پستان و امتیاز سلول‌های بدنی از دو گاوداری در اصفهان جمع‌آوری شد. ویرایش داده‌ها با نرم‌افزار SQL Server (نسخه ۲۰۱۲)، مدل‌سازی برای پیش‌بینی ورم پستان با نرم‌افزار WEKA (نسخه ۳/۸)، انجام شد. براساس نتایج به‌دست‌آمده، بهترین عملکرد مربوط به الگوریتم جنگل تصادفی در حالت کم‌نمونه‌برداری با صحت، حساسیت، تشخیص و ناحیه زیرمنحنی خم به‌ترتیب ۸۴/۳۰ درصد، ۹۴/۸۰ درصد، ۷۳/۸۰ درصد و ۰/۹۰ بود. بدون نمونه‌برداری، قدرت تشخیص موارد بیمار (حساسیت برحسب درصد) در الگوریتم‌های جنگل تصادفی، درخت تصمیم، بیز ساده و رگرسیون لجستیک به‌ترتیب ۱/۶۷، صفر، ۱۲/۲۹ و ۲/۰۶ بود که نسبت به استفاده از نمونه‌برداری به‌طور چشم‌گیری ضعیف‌تر بود. این به‌خاطر نامتوازن بودن تعداد موارد دو کلاس سالم و بیمار و نشان‌دهنده لزوم استفاده از روش‌های نمونه‌برداری بود. با توجه به یافته‌ها، الگوریتم درخت تصمیم نیز در روش کم‌نمونه‌برداری با اختلاف کمی بعد از جنگل تصادفی بهترین عملکرد را با صحت، حساسیت، تشخیص و ناحیه زیرمنحنی خم به‌ترتیب ۸۴/۰ درصد، ۹۴/۲ درصد، ۷۳/۹ درصد و ۰/۹۰ داشت. با توجه به هزینه‌ی محاسباتی بسیار بیش‌تر جنگل تصادفی نسبت به درخت تصادفی، در مواقعی که حجم داده‌ها بالاست، بهتر است از درخت تصمیم استفاده شود.

استناد: فیاضی‌کیا، محمدتقی؛ دادپسند، محمد و کشاورزی، حمیده (۱۴۰۲). استفاده از الگوریتم‌های یادگیری ماشین برای پیش‌بینی وقوع ورم پستان بالینی در گاوهای هلشتاین. *نشریه تولیدات دامی*، ۲۵ (۲)، ۱۲۳-۱۳۲. DOI: https://doi.org/10.22059/jap.2023.349388.623708

ناشر: مؤسسه انتشارات دانشگاه تهران.

۱. مقدمه

با توجه به پیامدهای زیان‌آور ورم پستان بر رفاه گاو شیری و تولیدات دامی و در نتیجه زیان‌های اقتصادی حاصل از آن، می‌توان این بیماری را یکی از پرهزینه‌ترین بیماری‌های گاو شیری دانست (Azooz *et al*., 2020). ازجمله‌ی این زیان‌ها می‌توان به کاهش تولید و کیفیت شیر، کاهش عملکرد تولیدمثلی، هزینه‌های تشخیص و درمان گاوهای مبتلا و هم‌چنین افزایش حذف زودهنگام دام‌ها اشاره کرد (Jamali *et al*., 2018; Puerto *et al*., 2021). با توجه به اهمیت این بیماری، پژوهش‌های زیادی برای پیش‌بینی، تشخیص زودهنگام و افزایش بهبود ژنتیکی دام انجام شده است. بیش‌تر پژوهش‌ها حول دو زمینه پیش‌گیری و درمان ورم پستان قرار دارند که ارجحیت بر پیش‌گیری این بیماری است، چراکه پیش‌گیری علاوه بر کاهش پیامدهای ذکرشده، از اثر منفی آنتی‌بیوتیک‌ها بر تولیدات دامی در درازمدت نیز جلوگیری می‌کند (Cheng and Han, 2020). در حوزه پیشگیری ورم پستان، مدیریت بهداشتی و پیش‌بینی ورم پستان نقش اساسی را بازی می‌کنند. به‌دلیل تحت کنترل‌نبودن عوامل محیطی و هزینه‌های زیاد کنترل بهداشت دام‌پروری برای پرورش‌دهندگان، توانایی مدیریت بهداشتی برای پیش‌گیری ورم پستان محدود است (Azooz *et al*., 2020; Xiong *et al*., 2020). یکی از راه‌کارهای پیشگیری از ورم پستان، پیش‌بینی وقوع آن با استفاده از عوامل دخیل در وقوع این بیماری است. از پرکاربردترین شاخص‌ها می‌توان به شمار سلول‌های بدنی شیر اشاره کرد که به‌عنوان مرتبط ترین شاخص با ورم پستان شناخته می‌شود (Cheng and Han, 2020). از دیگر شاخص‌های مرتبط با ورم پستان می‌توان به سطح تولید شیر، فصل زایش، نوبت زایش و سابقه‌ی سقط اشاره کرد (Jamali *et al*., 2018; Keshavarzi *et al*., 2019; Mishra, 2017).

روش‌های متعددی به‌طور عمده بر پایه رگرسیون لجستیک برای پیش‌بینی احتمال وقوع ورم پستان استفاده شده است. یادگیری ماشین یکی از مناسب‌ترین روش‌ها برای پیش‌بینی انواع بیماری‌ها محسوب می‌شود و در حال حاضر کاربرد آن در حوزه علوم دامی، به‌شدت در حال گسترش است (Neethirajan, 2020; Hyde *et al*., 2020; Garcia *et al*., 2020; Abdul Ghafoor and Sitkowska, 2021). با توجه به پیشرفت روزافزون تکنولوژی در مزارع دام‌پروری و ثبت دائم وضعیت گاوهای شیری با حس‌گرهای مختلف و ثبت داده‌های زیاد با شاخصه‌های بسیار زیاد، یادگیری ماشین می‌تواند به بهترین وجه از این داده‌ها برای پیش‌بینی بیماری‌های دامی استفاده کند (Garcia *et al*., 2020). پیش‌بینی خودکار الگوی آلودگی به ورم پستان با استفاده از الگوریتم‌های یادگیری ماشین، امکان‌سنجی و پیشنهاد شد تا تشخیص سریع و پیش‌گیری یا درمان موارد در سطح گله فراهم شود (Garcia *et al*., 2020). مشکلی که برای مدل‌سازی مناسب از داده‌های به‌دست‌آمده از گاوداری‌ها وجود دارد، عدم توزیع متعادل در نسبت افراد مبتلا به ورم پستان به افراد سالم در گله است که پیش‌بینی افراد مبتلا را سخت می‌کند و مدل‌ها نمی‌توانند بهترین عملکرد خود را نشان دهند (Mishra, 2017). روش‌های نمونه‌برداری با متوازن‌ساختن داده‌ها یکی از تکنیک‌های غلبه بر این مشکل است. ازجمله این روش‌ها می‌توان به کم‌نمونه‌برداری و بیش‌نمونه‌برداری اشاره کرد (Mishra, 2017). در روش کم‌نمونه‌برداری داده‌های کلاس بزرگ‌تر حذف شده تا به‌اندازه کلاس کوچک‌تر برسد، امادر روش بیش‌نمونه‌برداری این کلاس کوچک‌تر است که با روش رونویسی یا میانگین‌گیری همسایه‌ها (Synthetic minority oversampling technique: SMOTE) افزایش یافته تا به‌اندازه‌ی کلاس بزرگ‌تر شود (Lin *et al*., 2017).

پیش‌بینی احتمال ورم پستان براساس عوامل دخیل در احتمال وقوع آن می‌تواند به کاهش پیامدهای این بیماری منجر شده و به بهبود رفاه دام و سودآوری گله کمک کند. هدف این پژوهش، پیش‌بینی احتمال وقوع ورم پستان براساس عوامل سطح گله–دام با استفاده از الگوریتم‌های یادگیری ماشین بود.

۲. روش‌شناسی پژوهش

۲–۱. داده و متغیر

برای پیش‌بینی ورم پستان در گاوهای هلشتاین، از اطلاعات زایش و بیماری دو گله گاو شیری واقع در شهر اصفهان طی سال‌های ۱۳۸۶ تا ۱۳۹۶ استفاده شد. از هر گاوداری داده‌های مربوط به سابقه بیماری ورم پستان، تولید شیر، امتیاز سلول‌های بدنی، اطلاعات مربوط به زایش شامل فصل و نوبت زایش جمع‌آوری شدند. داده‌ها با نرم‌افزار SQL Server (نسخه ۲۰۱۲) و R (نسخه ۳/۶) ویرایش شدند.

طبق رکوردهای درمانی، موارد مبتلا به ورم پستان گاوهایی بودند که تحت درمان قرار گرفته بودند و بقیه به‌عنوان موارد سالم ثبت شدند. درواقع در این پژوهش، فقط موارد ورم پستان بالینی در نظر گرفته شدند. بر این اساس روزهای شیردهی به دوره‌های یک‌ماهه تقسیم‌بندی شده و در هر ماه اگر گاوی تحت درمان ورم پستان قرار گرفته بود کد یک و در غیر این‌صورت کد صفر را دریافت می‌کرد. سابقه بیماری ورم پستان به‌ازای داشتن (کد یک) یا نداشتن (کد صفر) سابقه‌ی ورم پستان در طی ماه قبل در همان دوره دسته‌بندی شد. دیگر متغیرهای مستقلی که در ارتباط با ورم پستان (متغیر وابسته) انتخاب شدند، شامل نوبت زایش (یک تا پنج)، سطح تولید شیر روزانه تصحیح شده برای چربی (سه سطح کم تولید (کم‌تر از ۲۴/۱۴ کیلوگرم شیر در روز)، متوسط (بین ۲۴/۱۴ تا ۴۶/۲۲ کیلوگرم شیر در روز) و پرتولید (بیش‌تر از ۴۶/۲۲ کیلوگرم شیر در روز) محاسبه شده براساس میانگین (۳۵/۱۸ کیلوگرم) و انحراف معیار (۱۱/۰۴ کیلوگرم) تولید شیر در سطح گله، فصل زایش (یک (بهار) تا چهار (زمستان)) و امتیاز سلول‌های بدنی (تبدیل لگاریتمی شمارش سلول‌های بدنی ($SCS = log_2^{(\frac{SCC}{100000})} + 3$) ماهانه) بودند. با توجه به این‌که سابقه ورم پستان در ماه قبل برای اولین ماه شیردهی موجود نبود، این ماه از محاسبات حذف شد و موارد مبتلا به‌عنوان سابقه ورم برای ماه بعد در نظر گرفته شدند. بعد از ویرایش داده‌ها، درمجموع ۳۳۲۹۹۰ رکورد ماهانه به‌دست آمد که از بین آن‌ها ۳۲۱۶۰۱ مورد سالم (۹۶/۵۸ درصد) و ۱۱۳۸۹ (۳/۴۲ درصد) مورد بیمار بودند. جدول (۱) خلاصه آماری متغیرهای مورداستفاده برای پیش‌بینی احتمال وقوع ورم پستان را نشان می‌دهد.

جدول ۱. اطلاعات مربوط به متغیرهای استفاده‌شده برای پیش‌بینی ورم پستان در دو گله گاو شیری هلشتاین

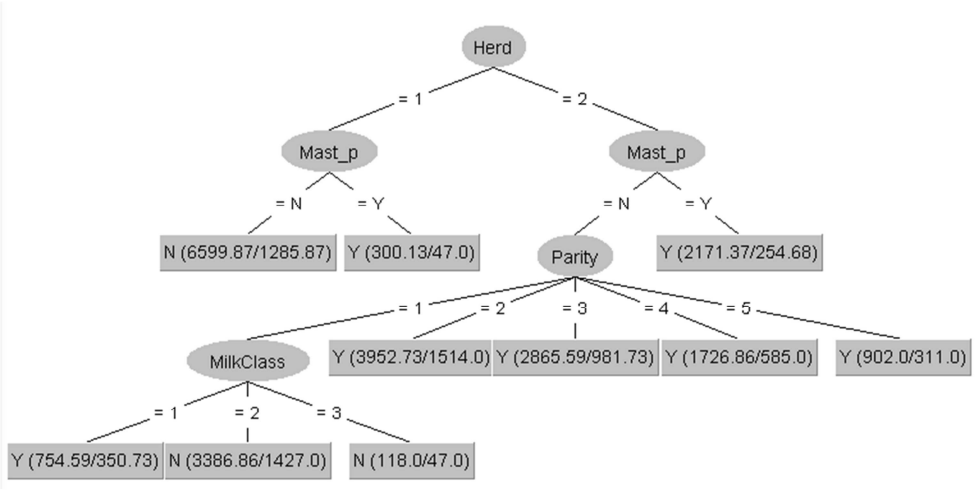
متغیرها	نوع داده	محدوده متغیر	نحوه محاسبه هر واحد
نوبت زایش	اسمی (nominal)	۱–۵	زایش منجر به افزایش نوبت شیردهی
فصل زایش	اسمی (nominal)	۱–۴	هر فصل یک واحد
سابقه ورم پستان	دودویی (binary)	۰/۱	وجود یا نبود سابقه ورم پستان در ماه قبل
مرحله شیردهی	اسمی (nominal)	۱–۱۰	هر ۳۰ روز یک مرحله
سطح تولید شیر	اسمی (nominal)	۱–۳	سطح ۱ (تولید شیر کم‌تر از ۲۴/۱۴ کیلوگرم در روز)، ۲ (بین ۲۴/۱۴ تا ۴۶/۲۲ کیلوگرم در روز) و ۳ (بیش‌تر از /۴۶ کیلوگرم در روز)
امتیاز سلول‌های بدنی	اسمی (nominal)	۱–۳	سطح ۱ یا پایین (SCS کم‌تر از ۱)، ۲ یا متوسط (بین یک تا ۲/۵) و ۳ یا بالا (بالاتر از ۲/۵)

۲–۲. الگوریتم‌های یادگیری ماشین

در این پژوهش، عملکرد چهار الگوریتم یادگیری ماشین با استفاده از نرم‌افزار WEKA (نسخه ۳/۸) (Markov and Russell, 2006) بررسی شدند. الگوریتم‌های مورداستفاده شامل بیز ساده، رگرسیون لجستیک، درخت تصمیم و جنگل تصادفی بودند که در ادامه مختصری از ویژگی‌های این الگوریتم‌ها آمده است.

درخت تصمیم: مدل‌های درختی شکلی (شکل ۱) هستند که در آن، آزمون ارزش‌های شاخصه در گره‌ها و برچسب‌های

کلاس در برگ‌ها انجام می‌شود. درخت‌های تصمیم نسبت به داده‌های پرت قوی و قادر به آموزش دسته‌بندها هستند، اما با وجود سادگی و درک آسان آن‌ها، چالش‌هایی ازجمله تعیین عمق بهینه درخت، انتخاب یک روش گزینش، متغیرهای مناسب و دست‌کاری داده‌های با مشاهده ازدست‌رفته وجود دارد (Charbuty and Abdulazeez, 2021).



شکل ۱. شکل شماتیک درخت تصمیم مورداستفاده در این پژوهش

جنگل تصادفی: یکی دیگر از الگوریتم‌های درختی است که در آن چندین درخت تصمیم که هرکدام شامل یک نمونه تصادفی از داده‌های ورودی است، در طی فرایند یادگیری تولیدشده و سپس با استفاده از سیستم رأی‌گیری خاصی، بهترین مدل برای داده‌ها انتخاب می‌شود. این الگوریتم نسبت به درخت تصمیم دقیق‌تر عمل می‌کند و یکی از مزیت های مهم این الگوریتم این است که برای برآورد ارزش‌های گمشده خیلی کارآمد بوده و می‌تواند وقتی نسبت زیادی از داده‌ها گمشده باشند، صحت بالای برآورد را حفظ کند (Liu and Zhang, 2012).

رگرسیون لجستیک: الگوریتمی بر پایه معادلات رگرسیون چندگانه و تابع لجستیک است. این الگوریتم برای پیش‌بینی مواردی که فقط دو کلاس دارند و درنتیجه فرض استقلال و توزیع استاندارد خطا مطرح نیست، به‌کار می‌رود (Dreiseitl and Ohno-Machado, 2020).

بیز ساده: الگوریتم بیز ساده بر پایه فرض عدم وابستگی صفات بنا نهاده شده است و به همین دلیل، انعطاف بالایی در حل مسائل مختلف دارد. تنها وقتی‌که وابستگی صفات بالا باشد، عملکرد این الگوریتم کاهش می‌یابد (Webb, 2010).

به‌دلیل نامتوازن‌بودن دسته‌ی بیمار (۳/۴۲ درصد) نسبت به‌دسته سالم (۹۶/۵۸ درصد) و عملکرد بهتر الگوریتم‌های یادگیری ماشین با داده‌های متعادل، از تکنیک‌های نمونه‌گیری استفاده شد (Shook *et al*., 2017; Rendon *et al*., 2020). دو روش کم‌نمونه‌برداری و بیش‌نمونه‌برداری برای متوازن‌ساختن دو دسته بیمار و سالم استفاده شد. در روش کم‌نمونه‌برداری برای متوازن‌ساختن داده‌های کلاس بزرگ‌تر به‌صورت تصادفی به‌اندازه کلاس کوچک‌تر کاهش می‌یابد و سپس الگوریتم‌های یادگیری ماشین برای ساختن مدل اجرا می‌شوند. برای بیش‌نمونه‌برداری از روش SMOTE استفاده شد که در آن داده‌های کلاس کوچک‌تر با در نظرگرفتن روش K-Nearest Neighbors (KNN) شبیه‌سازی شده تا به اندازه کلاس بزرگ‌تر برسند. برای کم‌نمونه‌برداری و بیش‌نمونه‌برداری از الگوریتم‌های SpreedSubSample و SMOTE در نرم‌افزار WEKA استفاده شد.

در برنامه WEKA برای تقسیم داده‌ها و ارزیابی عملکرد الگوریتم از روش 10-fold cross validation استفاده شد (Zigo *et al*., 2021; Kothoff *et al*.; 2019). در این روش، داده‌ها به‌صورت تصادفی به ۱۰ دسته تقسیم می‌شوند و سپس در ۱۰ نوبت هر بار نُه دسته به‌عنوان داده‌ی آموزشی در نظر گرفته می‌شوند و الگوریتم‌ها برای یافتن مدل مناسب بر آن اجرا شدند و یک دسته نیز برای ارزیابی مدل به‌دست‌آمده از آن دسته آموزشی در نظر گرفته شد. بعد از اجرای این روش ۱۰ مدل با ۱۰ عملکرد مختلف به‌دست آمد که برنامه WEKA میانگین آن‌ها را برای اعلام نتیجه عملکرد الگوریتم به‌کار می‌برد. برای ارزیابی عملکرد الگوریتم‌های مختلف از معیارهای مختلفی ازجمله صحت، حساسیت، تشخیص و AUC (Area under ROC curve) استفاده شد.

معیار صحت، نسبت موارد صحیح پیش‌بینی‌شده به‌کل موارد است که قدرت کلی پیش‌بینی با مدل‌های به‌دست‌آمده را نشان می‌دهد. حساسیت، نسبت موارد کلاس مثبت یا بیمار درست پیش‌بینی‌شده، به‌کل موارد بیمار است که می‌تواند نشانگر قدرت مدل در پیش‌بینی موارد بیمار باشد. تشخیص نیز قدرت مدل در پیش‌بینی موارد سالم است که از نسبت موارد سالم درست پیش‌بینی شده به کل موارد سالم به‌دست می‌آید. AUC نیز نسبت نرخ موارد مثبت درست پیش‌بینی‌شده به نرخ موارد منفی که به اشتباه به‌عنوان مثبت دسته‌بندی‌شده را نشان می‌دهد که بر این اساس می‌توان نتیجه گرفت که قدرت مدل در تفکیک موارد سالم از بیمار را نشان می‌دهد که بین صفر تا یک متغیر است. هرچه به یک نزدیک‌تر باشد، نشان‌دهنده درصد پیش‌بینی صحیح بالاتر و در نتیجه قدرت تفکیک بالاتر مدل است و هر چه به صفر نزدیک‌تر باشد پیش‌بینی‌ها اشتباه‌تر بوده است. البته با توجه به این‌که مقدار AUC ۰/۵ به معنی شانسی‌بودن پیش‌بینی و عدم تمایز بین دو کلاس محسوب می‌شود مقادیر کم‌تر از ۰/۵ به‌عنوان مقادیر غیرقابل قبول در نظر گرفته می‌شوند.

۳. یافته‌های پژوهش

با بررسی نتایج عملکرد الگوریتم‌ها در دو حالت استفاده از تکنیک کم‌نمونه‌برداری و تکنیک بیش‌نمونه‌برداری مشخص شد که در حالت استفاده از روش کم‌نمونه‌برداری (جدول ۲)، مقادیر صحت و حساسیت برای مدل‌هایی که با استفاده از الگوریتم‌های جنگل تصادفی، درخت تصمیم، بیز ساده و رگرسیون لجستیک تشکیل شدند، کم‌تر از حالت استفاده از روش بیش‌نمونه‌برداری بودند، امابرای حساسیت، عملکرد بیش‌تری نسبت به حالت بیش‌نمونه‌برداری داشتند.

جدول ۲. نتایج عملکرد چهار الگوریتم یادگیری ماشین مورداستفاده برای پیش‌بینی وقوع ورم پستان در گاوهای هلشتاین

الگوریتم	معیار ارزیابی عملکرد		
	صحت [۱]	حساسیت [۲]	تشخیص [۳]
روش کم‌نمونه‌برداری			
بیز ساده	۷۱/۱۰	۷۴/۷۰	۶۷/۴۰
رگرسیون لجستیک	۷۰/۷۹	۷۴/۴۰	۶۷/۱۰
درخت تصمیم	۸۴/۰۰	۹۴/۲۰	۷۳/۹۰
جنگل تصادفی	۸۴/۳۰	۹۴/۸۰	۷۳/۸۰
روش بیش‌نمونه‌برداری			
بیز ساده	۷۰/۱۴	۷۴/۸۰	۶۵/۵۰
رگرسیون لجستیک	۷۰/۸۰	۷۵/۹۰	۶۵/۷۰
درخت تصمیم	۷۵/۹	۸۴/۰۰	۶۷/۷۰
جنگل تصادفی	۷۶/۰۰	۸۴/۰۰	/۶۸۰۰

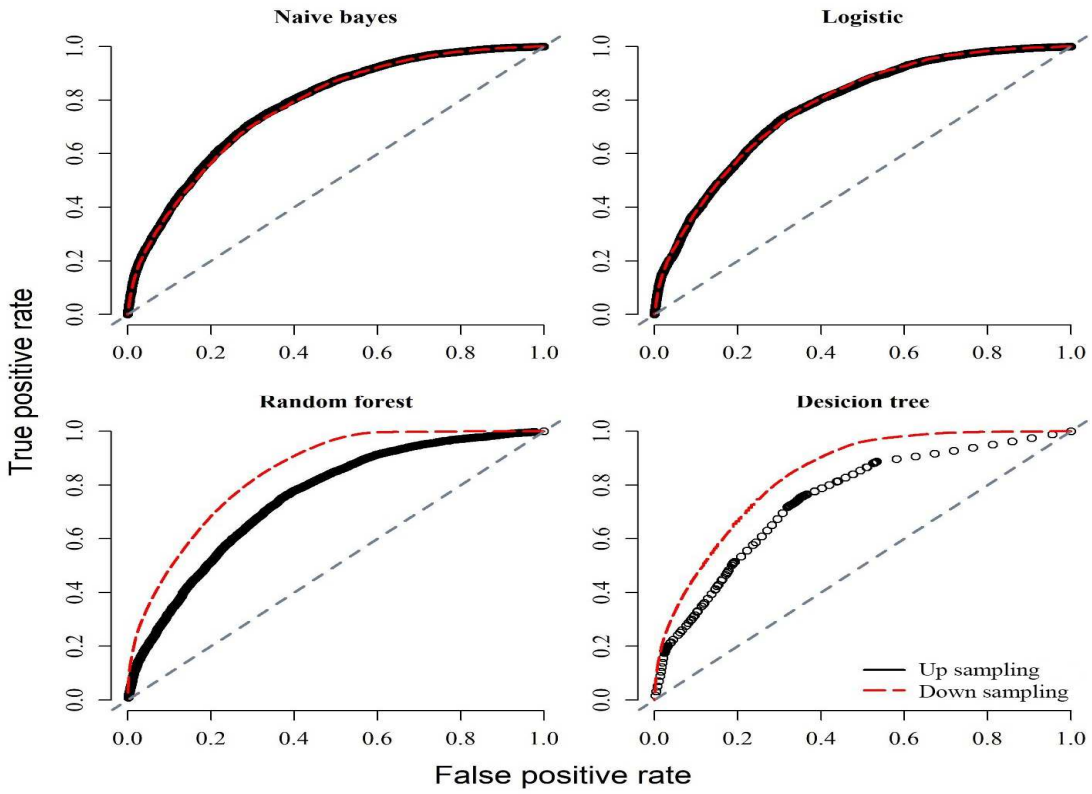
۱. صحت (Accuracy)، نسبت موارد صحیح پیش‌بینی شده به‌کل موارد است که قدرت کلی پیش‌بینی با مدل‌های به‌دست‌آمده را نشان می‌دهد.

۲. حساسیت (Sensitivity)، نسبت موارد کلاس مثبت یا بیمار درست پیش‌بینی‌شده، به‌کل موارد بیمار که می‌تواند نشانگر قدرت مدل در پیش‌بینی موارد بیمار باشد.

۳. تشخیص (Specifieity)، قدرت مدل در پیش‌بینی موارد سالم است که از نسبت موارد سالم درست پیش‌بینی‌شده به‌کل موارد سالم به‌دست می‌آید.

مقادیر AUC برای الگوریتم‌های بیز ساده، رگرسیون لجستیک، درخت تصمیم و جنگل تصادفی در حالت کم‌نمونه‌برداری به‌ترتیب ۰/۷۷، ۰/۷۸، ۰/۹۰ و ۰/۹۰ و برای حالت بیش‌نمونه‌برداری به‌ترتیب ۰/۷۷، ۰/۷۸، ۰/۸۴ و ۰/۸۵ بودند. به‌طور کلی عملکرد دو الگوریتم بیز ساده و رگرسیون لجستیک در هر دو حالت نمونه‌برداری نزدیک به یکدیگر بود و عملکرد دو الگوریتم درخت تصمیم و جنگل تصادفی به شکل قابل‌توجهی بهتر از دو الگوریتم دیگر بود (جدول ۲).

با توجه به این‌که در بین این معیارها حساسیت، به‌خاطر نشان‌دادن قدرت تشخیص موارد بیمار در مدل و در درجه بعدی AUC به‌خاطر نشان‌دادن قدرت تفکیک دو کلاس در مدل، از اهمیت بالاتری برخوردارند و با توجه به نتایج به دست‌آمده از عملکرد مدل‌ها، می‌توان نتیجه گرفت که در این پژوهش عملکرد مدل‌های رگرسیون لجستیک و بیز ساده در حالت استفاده از بیش‌نمونه‌برداری کمی بالاتر از حالت کم‌نمونه‌برداری بود (جدول ۲). با توجه به برتری حالت بیش‌نمونه‌برداری نسبت به بقیه حالت‌ها، اگر نمودار AUC مدل‌های به‌دست‌آمده در این حالت را در نظر بگیریم (شکل ۲)، مشخص است که الگوریتم جنگل تصادفی و درخت تصمیم در حالت کم‌نمونه‌برداری بالاترین میزان قدرت تفکیک دو کلاس بیمار و سالم را داشتند.



شکل ۲. نمودار خم **ROC** برای نشان دادن عملکرد چهار الگوریتم درخت تصمیم، رگرسیون لجستیک، بیز ساده و جنگل تصادفی در پیش‌بینی بیماری ورم پستان در گاوهای هلشتاین

۴. بحث

در پژوهشی که به پیش‌بینی ورم پستان در دوره نخست شیردهی، در گاوهای هلشتاین آمریکا با استفاده از دو الگوریتم بیز ساده و جنگل تصادفی پرداخته شد، الگوریتم جنگل تصادفی با صحت، تشخیص و AUC به‌ترتیب ۶۱ درصد، ۶۰/۲

درصد و ۶۵/۶ درصد، همانند پژوهش حاضر بالاترین عملکرد را داشت (Fadul-Pacheco *et al*., 2021). لازم به ذکر است که در پژوهش حاضر به‌دلیل نبود سابقه ورم پستان در تلیسه‌ها این معیار برای آن‌ها تهی در نظر گرفته شد، اما در این پژوهش فرایند پیش‌بینی برای تلیسه‌ها به‌صورت جداگانه در نظر گرفته شد و گزارش شد این پیش‌بینی جداگانه، در کوتاه‌مدت و میان‌مدت می‌تواند کارایی بالاتری برای پیش‌بینی ورم پستان داشته باشد؛ یک مدل پیش‌بینی برای گاوهایی که در دوره نخست شیردهی خطر ابتلای بیش‌تری دارند و مدل دیگر برای پیش‌بینی و مشخص‌کردن انفرادی گاوهایی که روزانه خطر ورم پستان در آن‌ها بالاست، به‌دست آمد. طبق پژوهشی در انگلستان، همانند مطالعه حاضر الگوریتم جنگل تصادفی به‌عنوان بهترین الگوریتم انتخاب شد که برای پیش بینی ورم پستان واگیردار و محیطی عملکرد معیارهای صحت، حساسیت و تشخیص به‌ترتیب ۹۸، ۸۶ و ۹۹ درصد برای حالت واگیردار و ۷۸، ۷۶ و ۸۱ درصد برای حالت محیطی گزارش شد (Hyde *et al*., 2020). هم‌چنین، در این پژوهش برخلاف پژوهش حاضر استفاده از روش‌های نمونه‌برداری تأثیری در افزایش عملکرد الگوریتم‌ها نداشت که احتمالاً به‌دلیل توازن داده‌ها در دو دسته بیمار و سالم بوده است.

در پژوهشی دیگر برای پیش‌بینی ورم پستان در گاوهای هلشتاین نیوزیلند از شش الگوریتم متفاوت که شامل بیز ساده و رگرسیون لجستیک، درخت تصمیم، جنگل تصادفی، مدل خطی تعمیم یافته و Gradient-Boosted Tree بودند، استفاده شد. از بین این دسته‌بندها، مدل به‌دست‌آمده با دسته‌بند Gradient-Boosted Tree که ترکیبی از الگوریتم درختی و رگرسیون است، بهترین عملکرد را به‌نمایش گذاشت و برخلاف پژوهش حاضر جنگل تصادفی پایین‌ترین عملکرد را داشت. صحت، حساسیت، تشخیص و AUC برای این مدل به‌ترتیب ۸۴ درصد، ۹۷ درصد، ۲۷ درصد و ۰/۸۱ بود (Ebrahimi *et al*., 2019). هم‌چنین، برخلاف پژوهش حاضر، روش‌های نمونه‌برداری موجب بهبود عملکرد مدل‌ها نشد. همین‌طور در پژوهشی دیگر با هدف پیش‌بینی ورم پستان و لنگش در گاوهای هلشتاین آلمان با استفاده از الگوریتم‌های رگرسیون لجستیک، Support Vector Machine، K-nearest neighbors، Gaussian Naïve Bayes، Extra Trees Classifier و جنگل تصادفی، مدل به‌دست‌آمده از الگوریتم Extra Trees Classifier که نسخه دیگر از الگوریتم جنگل تصادفی است، با AUC ۰/۷۹ برای ورم پستان و ۰/۷۱ برای لنگش، بالاترین عملکرد را نشان داد (Post *et al*., 2020). با این‌حال، با وجود عملکرد بالای جنگل تصادفی و Extra Trees Classifier عملکرد رگرسیون لجستیک هم بسیار نزدیک به این دو بود که بر همین اساس نویسنده نتیجه گرفته که می‌توان از الگوریتم‌های ساده مثل رگرسیون لجستیک نیز به‌جای الگوریتم‌های پیچیده استفاده کرد. قابل ذکر است که براساس الگوریتمهای جنگل تصادفی، نرم‌افزار کاربردی (masPA) طراحی شده است که می‌تواند خطر ورم پستان را با صحت ۹۸ درصد، حساسیت ۹۴ و تشخیص ۹۹ درصد پیش بینی کند که بر توانایی‌های یادگیری ماشین برای پیش بینی ورم پستان تاکید دارد (Abdul Ghafoor and Sitkowska, 2021).

۵. نتیجه‌گیری و پیشنهادها

مدل به‌دست‌آمده از الگوریتم جنگل تصادفی و کم‌نمونه‌برداری بهترین عملکرد را نشان داد و بعد از آن به‌ترتیب درخت تصمیم، رگرسیون لجستیک و بیز ساده قرار داشتند. البته، بیز ساده و رگرسیون لجستیک عملکرد نزدیکی نسبت به هم داشتند، امابا توجه به اهمیت بیش‌تر معیار حساسیت، عملکرد بیز ساده بالاتر بود. با وجود عملکرد بهتر الگوریتم جنگل تصادفی نسبت به درخت تصمیم، این دو الگوریتم، عملکرد نزدیکی نشان دادند و با توجه به هزینه محاسباتی بالای الگوریتم جنگل تصادفی نسبت به درخت تصمیم، در صورت حجم زیاد داده، الگوریتم درخت تصمیم پیشنهاد می‌شود. با

توجه به عملکرد بالای الگوریتم جنگل تصادفی با روش پیش‌پردازش SMOTE، در پیش‌بینی موارد بیمار، می‌توان آن را برای پیش‌بینی موارد مبتلا به ورم پستان در گاو شیری بخصوص در گله‌های با نرخ ابتلای بالا، پیشنهاد کرد.

۶. تشکر و قدردانی

از مدیران گله‌ها برای در اختیاردادن داده‌های لازم برای اجرای پژوهش، به‌طور ویژه تشکر و قدردانی می‌گردد.

۷. تعارض منافع

هیچ‌گونه تعارض منافع توسط نویسندگان وجود ندارد.

۸. منابع

Abdul Ghafoor, N., & Sitkowska, B. (2021). MasPA: A machine learning application to predict risk of mastitis in cattle from AMS sensor data. *AgriEngineering*, 3(3), 575-583.

Azooz, M. F., El-Wakeel, S. A., & Yousef, H. M. (2020). Financial and economic analyses of the impact of cattle mastitis on the profitability of Egyptian dairy farms. *Veterinary World*, 13(9), 1750-1759.

Bobbo, T., Biffani, S., Taccioli, C., Penasa, M., & Cassandro, M. (2021). Comparison of machine learning methods to predict udder health status based on somatic cell counts in dairy cows. *Scientific Reports*, 11(1), 1-10.

Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, 2(01), 20-28.

Cheng, W. N., & Han, S. G. (2020). Bovine mastitis: Risk factors, therapeutic strategies, and alternative treatments-A review. *Asian-Australasian Journal of Animal Sciences*, 33(11), 1699-1713.

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics*, 35(5-6), 352-359.

Ebrahimi, M., Mohammadi-Dehcheshmeh, M., Ebrahimie, E., & Petrovski, K. R. (2019). Comprehensive analysis of machine learning models for prediction of sub-clinical mastitis: Deep Learning and Gradient-Boosted Trees outperform other models. *Computers in biology and medicine*, 114, 103456.

Fadul-Pacheco, L., Delgado, H., & Cabrera, V. E. (2021). Exploring machine learning algorithms for early prediction of clinical mastitis. International Dairy Journal, 119, 105051-105060.

Garcia, R., Aguilar, J., Toro, M., Pinto, A., & Rodriguez, P. (2020). A systematic literature review on the use of machine learning in precision livestock farming. *Computers and Electronics in Agriculture*, 179, 105826-105838.

Hyde, R. M., Down, P. M., Bradley, A. J., Breen, J. E., Hudson, C., Leach, K. A., & Green, M. J. (2020). Automated prediction of mastitis infection patterns in dairy herds using machine learning. *Scientific reports*, 10(1), 1-8.

Jamali, H., Barkema, H. W., Jacques, M., Lavallée-Bourget, E. M., Malouin, F., Saini, V., ... & Dufour, S. (2018). Invited review: Incidence, risk factors, and effects of clinical mastitis recurrence in dairy cows. *Journal of dairy science*, 101(6), 4729-4746.

Keshavarzi, H., Sadeghi-Sefidmazgi, A., Stygar, A. H., & Kristensen, A. R. (2019). Abortion and other risk factors for mastitis in Iranian dairy herds. *Livestock Science*, 219, 40-44.

Kotthoff, L., Thornton, C., Hoos, H. H., Hutter, F., & Leyton-Brown, K. (2019). Auto-WEKA: Automatic model selection and hyperparameter optimization in WEKA. In Automated Machine Learning (pp. 81-95). Springer, Cham.

Lin, W. C., Tsai, C. F., Hu, Y. H., & Jhang, J. S. (2017). Clustering-based undersampling in class-imbalanced data. *Information Sciences*, 409, 17-26.

Liu, Y., Wang, Y., & Zhang, J. (2012). New machine learning algorithm: Random forest. In Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings 3 (pp. 246-252). Springer Berlin Heidelberg.

Markov, Z., & Russell, I. (2006). An introduction to the WEKA data mining system. *ACM SIGCSE Bulletin*, 38(3), 367-368.

Mishra, S. (2017). Handling imbalanced data: SMOTE vs. random undersampling. International *Research Journal of Engineering and Technology*, 4(8), 317-320.

Neethirajan, S. (2020). The role of sensors, big data and machine learning in modern animal farming. *Sensing and Bio-Sensing Research*, 29, 100367-100375.

Post, C., Rietz, C., Büscher, W., & Müller, U. (2020). Using sensor data to detect lameness and mastitis treatment events in dairy cows: A comparison of classification models. *Sensors*, 20(14), 3863.

Puerto, M. A., Shepley, E., Cue, R. I., Warner, D., Dubuc, J., & Vasseur, E. (2021). The hidden cost of disease: I. Impact of the first incidence of mastitis on production and economic indicators of primiparous dairy cows. *Journal of dairy science*, 104(7), 7932-7943.

Rendon, E., Alejo, R., Castorena, C., Isidro-Ortega, F. J., & Granda-Gutierrez, E. E. (2020). Data sampling methods to deal with the big data multi-class imbalance problem. *Applied Sciences*, 10(4), 1276-1291.

Shook, G. E., Kirk, R. B., Welcome, F. L., Schukken, Y. H., & Ruegg, P. L. (2017). Relationship between intramammary infection prevalence and somatic cell score in commercial dairy herds. *Journal of dairy science*, 100(12), 9691-9701.

Webb, G.I. (2010). Naïve Bayes. *Encyclopedia of Machine Learning*, 15, 713-714.

Xiong, Z., Cui, Y., Liu, Z., Zhao, Y., Hu, M., & Hu, J. (2020). Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation. *Computational Materials Science*, 171, 109203-109215.

Zigo, F., Vasil', M., Ondrašovičová, S., Výrostková, J., Bujok, J., & Pecka-Kielb, E. (2021). Maintaining optimal mammary gland health and prevention of mastitis. *Frontiers in veterinary science*, 8, 607311.