Shahid Bahonar University of Kerman

Agricultural Biotechnology Journal

Print ISSN: 2228-6705 Online ISSN: 2228-6500

Iranian Biotechnology Society

The role and diverse applications of machine learning in genetics, breeding, and biotechnology of livestock and poultry

Mohammadreza Mohammadabadi

*Corresponding Author. Professor, Animal Science Department, Faculty of Agriculture, Shahid Bahonar University of Kerman, Kerman, Iran. E-mail address: mrm@uk.ac.ir

Alireza Akhtarpoor

Department of Biology, Faculty of Science, Shahid Bahonar University of Kerman, Kerman, Iran. E-mail address: a.akhtarpoor@gmail.com

Amin Khezri

Associate Professor, Animal Science Department, Faculty of Agriculture, Shahid Bahonar University of Kerman, Kerman, Iran. E-mail address: akhezri@uk.ac.ir

Olena Babenko

Assistant Professor, Department of Animal Science, Bila Tserkva National Agrarian University, Bila Tserkva, Ukraine. E-mail address: lelya.babenko1978@gmail.com

Ruslana Volodymyrivna Stavetska

Assistant Professor, Department of Animal Science, Bila Tserkva National Agrarian University, Bila Tserkva, Ukraine. E-mail address: rstavetska@gmail.com

Iryna Tytarenko

Assistant Professor, Department of Animal Science, Bila Tserkva National Agrarian University, Bila Tserkva, Ukraine. E-mail address: iraponom80@ukr.net

Yulia Ievstafiieva

Associate professor, Department of Technologies of Livestock Production and processing, Higher Educational Institution “Podillia State University”, Ukraine. E-mail address: pp.nika22@ukr.net

Vita Buchkovska

Associate Professor, Department of Technologies of Livestock Production and processing, Higher Educational Institution “Podillia State University”, Ukraine. E-mail address: vbutschk@ukr.net

Viktor Slynko

Associate Professor, Poltava State Agrarian University, Ukraine. E-mail address: viktor.slynko@pdau.edu.ua

Volodymyr Afanasenko

Associate Professor, National University of Life and Environmental Sciences of Ukraine, Ukraine. E-mail address: afanasenko77@gmail.com

Abstract

Objective

Machine learning is a subset of artificial intelligence that is uniquely suited to address challenges in the fields of genetics, breeding, and biotechnology of livestock and poultry. By using algorithms that can learn patterns from data, machine learning enables accurate predictions, automated decision-making, and innovative solutions to complex problems in animal science. Unlike traditional statistical methods, which often assume linearity and independence among variables, machine learning is able to capture nonlinear relationships and interactions between genomic, environmental, and phenotypic factors. Therefore, the purpose of this study was to review the common types of machine learning algorithms used in livestock and poultry breeding, to outline their advantages and disadvantages, and to provide practical examples for these algorithms in the fields of genetics, breeding, and biotechnology of livestock and poultry.

Materials and Methods

In this study, by reviewing relevant databases and journals, studies related to machine learning in the field of genetics and breeding and biotechnology of livestock and poultry were searched using keywords. These studies were evaluated based on their design, methodology, results and relevance, and the main findings and concepts were extracted from them.

Results

The results showed that machine learning methods significantly outperform conventional methods. So that machine learning methods improve prediction accuracy and have smaller mean square error (MSE) and mean absolute error (MAE) in all scenarios. The findings also show the potential of combining classical bioinformatics methods with machine learning techniques to improve genomic prediction in the future. The results suggest machine learning algorithms as a promising tool to improve decision-making for livestock farmers. Machine learning analysis

improves monitoring methods and allows livestock farmers to identify animals that are likely to have problems in the future.

Conclusions

This study shows that the use of machine learning methods in the field of genetics, breeding, and biotechnology of livestock and poultry is increasing, and with this increase, the quality of machine learning methods used is also improving. Therefore, machine learning can play an important and prominent role in the sustainable development of livestock farming and provide benefits such as increased productivity in this field. Therefore, this study recommends that the use of machine learning methods and algorithms be promoted among the activists in the field of genetics, breeding, and biotechnology of livestock and poultry to identify and predict problems earlier and more accurately and prevent problems and economic losses.

Keywords: algorithm, bioinformatics, prediction, animal, artificial intelligence

Paper Type: Review Paper.

Citation: Mohammadabadi MR, Akhtarpoor A, Khezri A, Babenko O, Stavetska RV, Tytarenko I, Ievstafiieva Y, Buchkovska V, Slynko V, Afanasenko V (2024) The role and diverse applications of machine learning in genetics, breeding, and biotechnology of livestock and poultry. *Agricultural Biotechnology Journal* 16(4), 413-442.

Agricultural Biotechnology Journal 16(4), 413-442. DOI: 10.22103/jab.2025.24662.1644

Received: October 11, 2024. Received in revised form: December 21, 2024.

Accepted: December 22, 2024. Published online: December 30, 2024.

Publisher: Faculty of Agriculture and Technology Institute of Plant Production, Shahid Bahonar University of Kerman-Iranian Biotechnology Society.

مجله بیوتکنولوژی کشاورزی
شاپا چاپی: ۶۷۰۵-۲۲۲۸ شاپا الکترونیکی: ۶۵۰۰-۲۲۲۸

نقش و کاربردهای متنوع یادگیری ماشین در ژنتیک و اصلاح نژاد و بیوتکنولوژی دام و طیور

محمدرضا محمدآبادی

*نویسنده مسئول: استاد بخش علوم دامی، دانشکده کشاورزی، دانشگاه شهید باهنر کرمان، ایران. ایمیل: mrm@uk.ac.ir

علیرضا اخترپور

گروه زیست شناسی، دانشکده علوم، دانشگاه شهید باهنر کرمان، کرمان، ایران. ایمیل: a.akhtarpoor@gmail.com

امین خضری

دانشیار بخش علوم دامی، دانشکده کشاورزی، دانشگاه شهید باهنر کرمان، ایران. ایمیل: akhezri@uk.ac.ir

اولنا بابنکو

استادیار، گروه علوم دامی، دانشگاه ملی کشاورزی بیلا تسرکوا، بیلا تسرکوا، اوکراین. ایمیل: lelya.babenko1978@gmail.com

روسلانا ولودیمیریونا استاوتسکا

استادیار، گروه علوم دامی، دانشگاه ملی کشاورزی بیلا تسرکوا، بیلا تسرکوا، اوکراین. ایمیل: rstavetska@gmail.com

ایرینا تیتارنکو

استادیار، گروه علوم دامی، دانشگاه ملی کشاورزی بیلا تسرکوا، بیلا تسرکوا، اوکراین. ایمیل: iraponom80@ukr.net

یولیا ایوستافیوا

دانشیار، گروه فناوری‌های تولید و فرآوری دام، دانشگاه دولتی پودیلیا، اوکراین. ایمیل: pp.nika22@ukr.net

ویتا بوچکوفسکا

دانشیار، گروه فناوری های تولید و فرآوری دام، دانشگاه دولتی پودیلیا، اوکراین. ایمیل: vbutschk@ukr.net

ویکتور اسلینکو

دانشیار دانشگاه کشاورزی دولتی پولتاوا، اوکراین. ایمیل: viktor.slynko@pdau.edu.ua

ولودیمیر آفاناسنکو

دانشیار دانشگاه ملی علوم زیستی و محیطی اوکراین، اوکراین. ایمیل: afanasenko77@gmail.com

تاریخ دریافت: ۱۴۰۳/۰۷/۲۰ تاریخ دریافت فایل اصلاح شده نهایی: ۱۴۰۳/۱۰/۰۱ تاریخ پذیرش: ۱۴۰۳/۱۰/۰۲

چکیده

هدف: یادگیری ماشین، زیرمجموعه‌ای از هوش مصنوعی است که به طور منحصر به فردی برای مقابله با چالش‌های حوزه ژنتیک و اصلاح نژاد و بیوتکنولوژی دام و طیور بسیار مناسب است. با استفاده از الگوریتم‌هایی که می‌توانند الگوها را از داده‌ها یاد بگیرند، یادگیری ماشین پیش‌بینی‌های دقیق، تصمیم‌گیری خودکار و راه‌حل‌های نوآورانه برای مسائل پیچیده در علوم حیوانات را امکان‌پذیر می‌کند. برخلاف روش‌های آماری سنتی، که اغلب خطی بودن و استقلال را در بین متغیرها فرض می‌کنند، یادگیری ماشین قادر است روابط و تعاملات غیرخطی بین عوامل ژنومی، محیطی و فنوتیپی را ثبت کند. لذا، هدف این پژوهش بررسی انواع رایج الگوریتم‌های یادگیری ماشین که در پرورش دام و طیور استفاده می‌شوند، بیان مزایا و معایب آن‌ها و نیز بیان مثال‌های کاربردی برای این الگوریتم‌ها در حوزه ژنتیک و اصلاح نژاد و بیوتکنولوژی دام و طیور بود.

مواد و روش‌ها: در این پژوهش با بررسی پایگاه‌های داده و مجلات مربوطه، مطالعات مربوط به یادگیری ماشین در حوزه ژنتیک و اصلاح نژاد و بیوتکنولوژی دام و طیور با استفاده از کلمات کلیدی جستجو شدند. این مطالعات بر اساس طراحی، روش شناسی، نتایج و ارتباط آنها ارزیابی شد و یافته ها و مفاهیم اساسی از آن‌ها استخراج گردید.

نتایج: نتایج نشان داد که روش‌های یادگیری ماشین به طور قابل‌توجهی از روش‌های مرسوم بهتر عمل می‌کنند. به طوری که روش‌های یادگیری ماشین دقت پیش‌بینی را بهبود می‌بخشند و میانگین مربعات خطا (MSE) و میانگین خطای مطلق (MAE) کوچک‌تری را در همه سناریوها به همراه دارند. یافته‌ها همچنین پتانسیل ترکیب روش‌های بیوانفورماتیک کلاسیک با تکنیک‌های یادگیری ماشین را برای بهبود پیش‌بینی ژنومی در آینده نشان می‌دهد. نتایج الگوریتم‌های یادگیری ماشین را به عنوان ابزاری امیدوارکننده برای بهبود تصمیم‌گیری برای دامپروران پیشنهاد می‌کند. تجزیه و تحلیل یادگیری ماشین روش‌های نظارت را بهبود می‌بخشد و به دامپروران می‌کند تا حیواناتی را که احتمالاً در آینده مشکلاتی داشته باشند از قبل شناسایی کنند.

نتیجه‌گیری: این مطالعه نشان می‌دهد که استفاده از روش‌های یادگیری ماشین در حوزه ژنتیک و اصلاح نژاد و بیوتکنولوژی دام و طیور رو به افزایش است و با این افزایش کیفیت روش‌های یادگیری ماشین استفاده شده نیز رو به بهبود است. لذا، یادگیری ماشین می‌تواند در توسعه پایدار دامپروری و ارائه مزایایی مانند افزایش بهره‌وری در این حوزه نقش مهم و پررنگی را ایفا کند. بنابراین، این پژوهش توصیه می‌کند که استفاده از روش‌ها و الگوریتم‌های یادگیری ماشین در بین فعالان حوزه ژنتیک و اصلاح نژاد و بیوتکنولوژی دام و طیور ترویج شود تا زودتر و دقیق‌تر مشکلات را شناسایی و پیش بینی کنند و از بروز مشکلات و ضررهای اقتصادی جلوگیری نمایند.

کلیدواژه‌ها: الگوریتم، بیوانفورماتیک، پیش‌بینی، حیوان، هوش مصنوعی

نوع مقاله: مروری.

استناد: محمدآبادی محمدرضا، اخترپور علیرضا، خضری امین، بابنکو اولنا، استاوتسکا روسلانا ولودیمیریونا، تیتارنکو ایرینا، ایوستافیوا یولیا، بوچکوفسکا ویتا، اسلینکو ویکتور، آفاناسنکو ولودیمیر (۱۴۰۳) نقش و کاربردهای متنوع یادگیری ماشین در ژنتیک و اصلاح نژاد و بیوتکنولوژی دام و طیور. مجله بیوتکنولوژی کشاورزی، ۱۶(۴)، ۴۱۳-۴۴۲.

Publisher: Faculty of Agriculture and Technology Institute of Plant Production, Shahid Bahonar University of Kerman-Iranian Biotechnology Society.

مقدمه

تقاضای فزاینده برای محصولات دام و طیور مستلزم اقدامات پایدار و کارآمد در تولیدات آن‌ها است. رویکردهای سنتی به پرورش دام و طیور و بیوتکنولوژی، در حالی که از بسیاری جهات موثر هستند، به طور فزاینده‌ای توسط پیچیدگی‌های سیستم‌های مدرن دام و طیور محدود می‌شوند. ظهور فن آوری‌های توالی‌یابی با توان بالا و انباشت مقادیر زیادی از داده‌های ژنتیکی و فنوتیپی فرصت‌های جدیدی را برای اصلاح برنامه‌های اصلاح نژادی و کاربردهای بیوتکنولوژی ایجاد کرده است. با این حال، حجم، پیچیدگی و چند بعدی بودن این مجموعه داده‌ها نیازمند ابزارهای تحلیلی پیشرفته است. یادگیری ماشین[1] (ML)، زیرمجموعه‌ای از هوش مصنوعی[2] (AI) است که به طور منحصر به فردی برای مقابله با این چالش‌ها بسیار مناسب است (Mohammadabadi et al. 2024). با استفاده از الگوریتم‌هایی که می‌توانند الگوها را از داده‌ها یاد بگیرند، MLپیش‌بینی‌های دقیق، تصمیم‌گیری خودکار و راه‌حل‌های نوآورانه برای مسائل پیچیده در علوم حیوانات را امکان‌پذیر می‌کند (Li et al. 2024). برخلاف روش‌های آماری سنتی، که اغلب خطی بودن و استقلال را در بین متغیرها فرض می‌کنند، ML قادر است روابط و تعاملات غیرخطی بین عوامل ژنومی، محیطی و فنوتیپی را ثبت کند. یکی از تاثیرگذارترین کاربردهای ML انتخاب ژنومی[3] (GS) است که در آن ارزش‌های اصلاحی بر اساس نشانگرهای ژنتیکی پیش‌بینی می‌شوند. روش‌های سنتی مانند بهترین پیش‌بینی نااریب خطی[4] (BLUP) توسط الگوریتم‌های ML، از جمله ماشین‌های بردار پشتیبان[5] (SVM)، جنگل‌های تصادفی[6] (RFs)، و تکنیک‌های یادگیری عمیق مانند شبکه‌های عصبی کانولوشنال جایگزین می‌شوند (Mohammadabadi et al. 2024). این الگوریتم‌ها دقت بالاتری را ارائه می‌دهند، به‌ویژه برای صفاتی که تحت تأثیر فعل و انفعالات پیچیده ژنتیکی و محیطی قرار دارند. یکی دیگر از حوزه های کلیدی فنوتیپ است، جایی که ML خودکارسازی ارزیابی صفات را فعال کرده است. فن‌آوری‌هایی مانند بینایی رایانه و حسگرهای پوشیدنی داده‌های بلادرنگ را در مورد سلامت، رفتار و عملکرد حیوانات جمع‌آوری می‌کنند، که الگوریتم‌های ML آن‌ها را برای شناسایی الگوها و ناهنجاری‌ها پردازش می‌کنند (Chafai et al. 2023). در بیوتکنولوژی حیوانی، ML ویرایش دقیق ژن را تسهیل می‌کند و ابزارهایی مانند CRISPR-Cas9را برای کاربردهای خاص بهینه می‌کند. به عنوان مثال، مدل‌های ML اثرات خارج از هدف را پیش‌بینی می‌کنند و کارایی RNA را هدایت می‌کنند و ایمنی و کارایی تغییرات ژنتیکی را افزایش می‌دهند (Zhang et al., 2020). این مطالعه به بررسی کاربردهای متنوع ML در ژنتیک دام و طیور، اصلاح نژاد و بیوتکنولوژی می‌پردازد و بر مثال‌های عملی، چالش‌ها و

[1] machine learning
[2] artificial intelligence
[3] genomic selection
[4] best linear unbiased prediction
[5] support vector machines
[6] random forests

چشم‌اندازهای آینده تأکید می‌کند. با ادغام ML با روش‌های سنتی، محققان و دامپروران می‌توانند فرصت‌های جدیدی را برای پیشرفت‌های پایدار در علوم حیوانات باز کنند.

یادگیری ماشین

یادگیری ماشین حوزه‌ای از هوش مصنوعی است که بر توسعه الگوریتم‌های رایانه‌ای تمرکز دارد که می‌توانند به طور خودکار عملکرد خود را در یک کار خاص از طریق تجربه بهبود بخشند. به عبارت دیگر، یادگیری ماشینی برنامه‌های کامپیوتری را قادر می‌سازد تا از داده‌ها یاد بگیرند و توانایی خود را برای پیش‌بینی یا تصمیم‌گیری بدون برنامه‌ریزی صریح بهبود بخشند. الگوریتم‌های یادگیری ماشین برای شناسایی الگوها در داده‌ها و استفاده از آن الگوها برای پیش بینی یا انجام اقدامات طراحی شده‌اند. این امر از طریق استفاده از روش‌های آماری، تکنیک‌های بهینه‌سازی و سایر روش‌های ریاضی به دست می‌آید که به رایانه اجازه می‌دهد از مقادیر زیادی داده یاد بگیرد. این امر در طیف گسترده‌ای از برنامه‌ها، مانند تشخیص تصویر، پردازش زبان طبیعی، تجزیه و تحلیل پیش بینی، و سیستم‌های توصیه‌ای استفاده می‌شود. همچنین یکی از اجزای حیاتی بسیاری از فناوری‌های پیشرفته، از جمله اتومبیل‌های خودران، پزشکی شخصی‌سازی‌شده و تشخیص گفتار است. الگوریتم‌های یادگیری ماشین را می‌توان در اصلاح نژاد دام و طیور برای پیش‌بینی و بهبود صفات مختلف مانند سرعت رشد، تولید شیر، تولید تخم مرغ، باروری، مقاومت در برابر بیماری‌ها و موارد دیگر مورد استفاده قرار داد. در ادامه برخی از انواع رایج الگوریتم‌های یادگیری ماشین که در پرورش دام و طیور استفاده می‌شوند به تفصیل خواهند آمد.

رگرسیون خطی

رگرسیون خطی یک الگوریتم ساده است که می‌تواند مقدار یک متغیر وابسته را بر اساس یک یا چند متغیر مستقل پیش بینی کند. در اصلاح نژاد دام و طیور می‌توان از رگرسیون خطی برای پیش بینی ارزش یک صفت خاص مانند تولید شیر یا سرعت رشد بر اساس عوامل مختلفی مانند سن، وزن و مصرف خوراک استفاده کرد (Gianola & de los Campos 2018). این رگرسیون مدلی است که معمولاً برای پیش‌بینی مقدار متغیر پیوسته y که برچسب یا متغیر هدف نیز نامیده می‌شود، با استفاده از اصطلاحات ML، از طریق بردار متغیرهای توضیحی که متغیرهای مستقل یا ویژگی‌های x نامیده می‌شوند و یک تابع خطی استفاده می‌شود. اگر مدل شامل یک متغیر مستقل x باشد، رگرسیون خطی ساده رابطه بین متغیرها را با استفاده از مدل زیر تعریف می‌کند:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

که در آن β0 اصطلاح قطع[7] و β1 یک ضریب رگرسیون است که نشان‌دهنده تغییر در نتیجه برای افزایش ۱ واحدی در مقدار متغیر مستقل x است و ε نشان‌دهنده اصطلاح خطا است که نویز نیز نامیده می‌شود. متغیر وابسته y را می توان با بیش از یک متغیر توضیحی توضیح داد. در این مورد، ما در مورد رگرسیون خطی چند متغیره[8] (MLR) صحبت می‌کنیم (Maulud & Abdulazeez 2020). مدل اصلی برای MLR به صورت زیر است:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m + \varepsilon$$

رگرسیون خطی یک الگوریتم یادگیری نظارت شده در نظر گرفته می‌شود زیرا ما مدل را با مجموعه داده‌ای حاوی ویژگی های x_i و مقادیر متناظر متغیر هدف y_i تغذیه می‌کنیم و انتظار داریم پیش بینی دقیق y_j برای مجموعه دیگری از ویژگی‌ها x_j باشد. به منظور دستیابی به دقت کافی، مدل مقدار یک تابع زیان انتخابی[9] را به حداقل می رساند. متداول‌ترین تابع زیان مورد استفاده برای رگرسیون خطی، خطای حداقل مربعات (LSE) است (Nasteski 2017). مدل‌های ML اغلب بسته به ویژگی‌های مجموعه داده‌هایی که روی آن‌ها اعمال می‌شوند، عملکرد متفاوتی از خود نشان می‌دهند. عوامل متعددی وجود دارد که بر چگونگی عملکرد یک مدل بر روی یک مجموعه داده خاص تأثیر می‌گذارد. برخی از مدل‌ها، مانند شبکه‌های یادگیری عمیق، با مجموعه داده‌های بزرگ بهتر عمل می‌کنند، در حالی که مدل‌های ساده‌تر مانند رگرسیون خطی با مجموعه داده‌های کوچک‌تر به خوبی کار می‌کنند. مجموعه داده‌هایی با ویژگی‌های ابعادی بالا یا انواع داده‌های مختلط (عددی، طبقه‌ای) ممکن است از مدل‌هایی بهره ببرند که می‌توانند چنین پیچیدگی‌هایی مانند ماشین‌های بردار پشتیبان (SVM) و مدل‌های مبتنی بر درخت را مدیریت کنند. مجموعه داده‌های با ابعاد بالا نیز ممکن است به تکنیک‌های کاهش ابعاد قبل از اعمال یک مدل ML نیاز داشته باشند. برخی از مدل‌ها نسبت به نویز و نقاط پرت قوی‌تر هستند (مانند درخت‌های تصمیم‌گیری، جنگل‌های تصادفی)، در حالی که برخی دیگر (مثلاً رگرسیون خطی) ممکن است حساس‌تر باشند و در حضور آنها ضعیف عمل کنند (Crossa et al. 2024).

رگرسیون لجستیک

رگرسیون لجستیک[10] یک مدل طبقه بندی است که به طور منظم برای تجزیه و تحلیل نتایج دوگانه یا دوتایی استفاده می‌شود (لاوالی، ۲۰۰۸) .به عبارت دیگر، رگرسیون لجستیک برای مطالعه اثرات متغیرهای پیش بینی کننده بر نتایج باینری یا مقوله‌ای، مانند وجود یا عدم وجود یک رویداد استفاده می‌شود (نیک و کمپبل، ۲۰۰۷) .داده‌های آموزشی به مدلی داده می‌شود که از یک تابع لجستیک برای پیش بینی احتمال رویداد استفاده می‌کند. بر خلاف رگرسیون خطی، رگرسیون لجستیک نیازی به رابطه خطی بین

[7] intercept term
[8] multivariate linear regression
[9] chosen loss function
[10] logistic regression

متغیرهای وابسته و مستقل ندارد، مدل از یک تبدیل لاگ به نسبت شانس تعریف شده به عنوان نسبت احتمال وقوع رویداد تقسیم بر احتمال رخ ندادن رویداد استفاده می‌کند (Nick & Campbell 2007; LaValley 2008). فرضیه رگرسیون لجستیک به صورت زیر تعریف می‌شود:

$$h_\theta(x) = g(\theta^T x)$$

که در آن تابع g یک تابع سیگموئیدی است که به صورت زیر تعریف می‌شود:

$$g(z) = \frac{1}{1 + e^{-z}}$$

رگرسیون لجستیک از یک تابع زیان برآورد حداکثر احتمال[11] (MLE) که یک احتمال شرطی است استفاده می‌کند. الگوریتم هر مشاهده را بر اساس بزرگتر یا کوچکتر بودن احتمال از یک آستانه معین، مثلا ۰/۵، به کلاس ۰ یا کلاس ۱ اختصاص می‌دهد (Belyadi & Haghighat 2021). داده های عدم تعادل دارای توزیع طبقاتی نامتعادل هستند و مدل‌هایی مانند رگرسیون لجستیک ممکن است بدون تنظیمات مناسب با مشکل مواجه شوند. مدل‌های پیچیده مانند شبکه‌های عمیق، به ویژه در مجموعه داده‌های کوچک یا پر سر و صدا مستعد بیش از حد برازش هستند. مدل‌های ساده تر، اگرچه قدرت کمتری دارند، ممکن است در چنین مواردی بهتر تعمیم دهند.

مثال کاربردی: اشکال جدی زیربنای حاشیه نویسی بیولوژیکی داده‌های توالی کل ژنوم، مسئله $p >> n$ است، به این معنی که تعداد واریانت‌های چندشکلی (p) بسیار بیشتر از تعداد رکوردهای فنوتیپی موجود (n) است. در پژوهشی Kotlarz et al. (2024) راهی برای دور زدن مشکل با ترکیب رگرسیون لجستیک LASSO با یادگیری عمیق برای طبقه‌بندی گاوها به عنوان حساس یا مقاوم به ورم پستان، بر اساس ژنوتیپ‌های پلی‌مورفیسم تک نوکلئوتیدی (SNP) پیشنهاد کردند. در این پژوهش تعداد ۵۲ گاو هلشتاین–فریزین از یک گله که در آن ۹۹۱ مورد بالینی ورم پستان مشاهده شده بود مطالعه شدند. این ۵۲ گاو به یک گروه آموزشی متشکل از ۳۲ گاو و یک گروه آزمایشی از ۲۰ گاو باقی مانده تقسیم شدند. در گروه آموزشی، گاوها خواهر و برادر ناتنی پدری بودند که از نظر تعداد باروری های ثبت شده، سطح تولید و سال تولد مطابقت داشتند، اما در وضعیت مقاومت آنها به ورم پستان تفاوت داشتند. بنابراین، ۱۶ گاو مقاوم به ورم پستان بودند و هیچ گونه بروز ورم پستان بالینی در طول عمر تولیدی خود نداشتند، در حالی که ۱۶ گاو حساس به ورم پستان بودند و تحت بروز چندین بیماری قرار داشتند. DNA ژنومی آنها با پلت فرم Illumina HiSeq2000در حالت جفتی با طول خواندن ۱۰۰ جفت باز تعیین توالی شد. تعداد خوانش‌های خام تولید شده برای یک حیوان از ۱۶۴۹۸۴۱۴۷ تا ۴۷۲۲۶۵۶۲۰ متغیر بود. گروه آزمایش شامل ۱۰ گاو حساس به ورم پستان و ۱۰ گاو مقاوم به ورم پستان بود که با پلت فرم Illumina NovaSeq 6000 با طول خواندن ۱۵۰ جفت باز تعیین توالی شدند.

[11] maximum likelihood estimation

مجموعه دیگری از گاوهای هلشتاین–فریزین با سوابق ورم پستان بالینی از پایگاه داده PLOWET برای صفات بهداشتی ثبت شده توسط دامپزشک در چهار مزرعه استفاده شد. در بین ۱۴۹۹ نفر، ورم پستان بالینی برای ۷۱۲ گاو ثبت شد. بیشتر گاوها با استفاده از Illumina BovineSNP50K تعیین ژنوتیپ شدند. علاوه بر این، سوابق شجره چند نسلی از سال ۱۹۱۴، شامل ۸۹۴۴ اجداد گاوهای ژنوتیپ شده، برای تخمین رابطه ژنتیکی افزایشی آنها در دسترس بود. آزمایشات بیوانفورماتیکی برای شناسایی اسنیپ‌ها نیز اجرا شد. اولین قدم برای غلبه بر مشکل $p >> n$ ، پیش انتخاب SNP با استفاده از مدل رگرسیون لجستیک زیر انجام شد:

$$P(y = 1|X) = \frac{e^{\beta^T X}}{1 + e^{\beta^T X}},$$

که در آن $P(y = 1|X)$ احتمال حساس بودن به ورم پستان را برای هر گاو مشروط به ژنوتیپ‌های SNP (X) با جریمه LASSO(λ) نشان می دهد که بر برآورد کننده اثر اسنیپ ($\hat{\beta}$) اعمال می‌شود:

$$\hat{\beta} = \text{argmin}_\beta \left\{ -\sum_{i=1}^{N_{cow}} \left[y_i \log\left(\beta^T x_i\right) + (1 - y_i) \log\left(1 - \beta^T x_i\right) \right] + \lambda \sum_{j=1}^{N_{SNP}} \left|\beta_j\right| \right\}$$

این مدل رگرسیون لجستیک با پایتون از طریق کتابخانه یادگیری Scikit (Pedregosa et al. 2011) با استفاده از روش بهینه‌سازی احتمال شیب افزایشی با پشتیبانی از اهداف ترکیبی غیر محدب[12] (Defazio et al. 2014) و تنظیم L1 برای تخمین اثر SNP پیاده‌سازی شد. جریمه به صورت $C = \frac{1}{\lambda}$ بیان شد، در حالی که جریمه های مختلفی با استفاده از یک شبکه در C در بازه $(0.1; 1.0]$ با گامه ۰/۱ اجرا شد. هرچه λ کوچکتر باشد، تخمین SNP بیشتری روی صفر تنظیم می‌شود.

در این پژوهش، از میان چندین معماری، معماری با ۲۰۴۶۴۲ اسنیپ به عنوان بهترین انتخاب شد. این معماری از دو لایه به ترتیب با ۷ و ۴۶ واحد در هر لایه تشکیل شده بود که نرخ خروج به ترتیب ۰/۲۱۰ و ۰/۳۵۸ را اجرا می‌کردند. طبقه بندی داده‌های آزمون به AUC[13] برابر با ۰/۷۵۰، دقت مساوی ۰/۶۵۰، حساسیت برابر با ۰/۶۰۰، و ویژگی مساوی ۰/۷۰۰ منجر شد. اسنیپ‌های معنی‌دار بر اساس Hapley Additive explanation (SHAP) انتخاب شدند. به عنوان یک نتیجه نهایی، یک عبارت هستی‌شناسی ژن[14] (GO) مربوط به فرآیند بیولوژیکی و سیزده اصطلاح GO مربوط به عملکرد مولکولی به طور قابل توجهی در مجموعه ژنی غنی شد که با اسنیپ‌های معنی‌دار مطابقت داشت. یافته‌های این پژوهش نشان داد که رویکرد بهینه می‌تواند به درستی

[12] incremental gradient likelihood optimization method with support for non-strongly convex composite objectives
[13] area under the curve
[14] gene ontology

حساسیت یا وضعیت مقاومت را برای تقریباً ۶۵ درصد گاوها پیش‌بینی کند. ژن هایی که با مهم ترین اسنیپ‌ها مشخص شده بودند به پاسخ ایمنی و سنتز پروتئین مربوط می شدند.

این پژوهشگران بر اهمیت بهره‌برداری از جنبه‌های متعدد تحلیل بیوانفورماتیک داده‌های بیولوژیکی، که فراتر از کاربرد نرم‌افزار بیوانفورماتیک است و همچنین شامل عناصر انتخاب ویژگی در داده‌های چند بعدی است که امروزه در ژنومیک، انتخاب مدل چند سطحی، تجزیه و تحلیل آماری و در نهایت تفسیر بیولوژیکی معمول است تأکید کردند. این رویکرد به ویژه در تجزیه و تحلیل فنوتیپ‌هایی با حالت وراثت پیچیده، مانند ورم پستان بالینی، که تحت تأثیر ژن‌های متعدد با اثرات متفاوت است و نه لزوماً تنها چند ژن با اثرات بزرگ، مهم است. علاوه بر این، به دلیل وراثت پذیری کم تا متوسط، تأثیر این ژن‌ها احتمالاً به محیط نیز وابسته است. علاوه بر این، آن‌ها نشان دادند که از آنجایی که در زیست شناسی، به دلیل محدودیت های مالی، اخلاقی یا در دسترس بودن داده ها، دستیابی به مجموعه داده‌های بسیار بزرگ همیشه امکان پذیر نیست، برنامه‌های کاربردی یادگیری ماشین باید به دقت بر روی انتخاب معماری مدل‌ها و فراپارامترهای آنها تمرکز کنند. در مورد اندازه محدود داده‌های ورودی، تنها چنین رویکرد انتخاب مدل گسترده اجازه می‌دهد تا دقت طبقه بندی معقول یا پیش بینی دقیق به دست آید.

درختان تصمیم

درختان تصمیم[15] (DT)، همچنین به عنوان درختان طبقه بندی و رگرسیون[16] (CART) شناخته می‌شوند، یکی از محبوب‌ترین الگوریتم‌های یادگیری نظارت شده بر اساس پارتیشن بندی بازگشتی[17] است (Jiang et al. 2020). این رویکرد اولین بار توسط Breiman et al. (1984) معرفی شد و متکی بر تقسیم یک مجموعه داده بزرگ ناهمگن به چندین زیر مجموعه کوچکتر همگن است که منجر به یک ساختار انشعابی می‌شود. این ساختار شامل گره‌هایی است که از طریق شاخه‌ها به هم متصل شده‌اند. اگر یک گره نشان دهنده یک یال ورودی نباشد، ریشه نامیده می‌شود. به طور کلی، همه گره‌ها دارای یک یال ورودی و دو یا چند یال خروجی هستند. گره‌هایی که لبه‌های خروجی ندارند برگ نامیده می‌شوند. در درخت‌های تصمیم، تقسیم داده‌های آموزشی با پاسخ دادن به چندین سوال به صورت تدریجی از بالاترین گره تا یک برگ انجام می‌شود. یک سوال خوب می‌تواند یک مجموعه داده ناهمگن را به چند نمونه فرعی همگن تقسیم کند. درختان تصمیم می‌توانند با مشکلات طبقه بندی و رگرسیون مقابله کنند. برای متغیرهای پیوسته، تقسیم با استفاده از یک آستانه انجام می‌شود، قانون به شکل $x < s$ است که در آن s یک آستانه بر روی متغیر x است. برعکس، وقتی متغیر گسسته است، تقسیم به شکل $x \in L$ است که در آن L زیر مجموعه‌ای از سطوح ممکن x است. هنگامی که متغیر هدف پیوسته است، به این معنی که ما با رگرسیون سر و کار داریم، مقدار پیش‌بینی‌شده هر زیرگروه، مقدار

[15] decision trees
[16] classification and regression trees
[17] recursive partitioning

متوسط y برای همه مشاهدات در مجموعه آموزشی اختصاص داده شده به آن زیر گروه است (Crisci et al. 2012). در مقابل، زمانی که y گسسته است و الگوریتم DT با مشکلات طبقه‌بندی سروکار دارد، بیشترین سطح y روی مشاهده برگ به مقدار هدف اختصاص می‌یابد. الگوریتم اصلی مورد استفاده برای ساخت درخت‌های تصمیم‌گیری برای موضوعات رگرسیون Iterative Dichotomiser 3 (ID3) است که از کاهش انحراف استاندارد[۱۸] (SDR) برای تولید درخت تصمیم استفاده می‌کند. در موقعیت‌های طبقه‌بندی، الگوریتم ID3 از آنتروپی استفاده می‌کند که به عنوان معیاری برای همگنی نمونه‌های فرعی و به دست آوردن اطلاعات تعریف می‌شود (Choudhary and Gianey 2017). این روش به دلیل انعطاف پذیری و سهولت تفسیر پذیری آن بسیار مورد استفاده قرار می‌گیرد.

مثال کاربردی: یک پژوهش تحت عنوان هوش مصنوعی و روش‌های کلاسیک در ژنتیک و اصلاح نژاد حیوانات توسط Soloshenkov et al. (2024) انجام شد. آن‌ها روش‌های اساسی ژنتیک جمعیت و پرورش حیوانات و روش‌های ریاضی یادگیری ماشین مورد استفاده در اصلاح نژاد حیوانات را تجزیه و تحلیل کردند. این پژوهشگران مدل‌های کتابخانه CatBoost را بر روی نمونه دو گونه اهلی اسب (Equus caballus) و گوزن شمالی (Rangifer tarandus) آموزش دادند. آن‌ها داده‌های پانل‌های ریزماهواره مکان‌های ۱۶ و ۱۷ به ترتیب برای آموزش مدل با استفاده از داده‌های گوزن شمالی، نژادهای اسب اروپایی و روسی اهلی و وحشی را استفاده کردند. شاخص‌های استاندارد (دقت، صحت، یادآوری و F1)را محاسبه کردند و ماتریس‌های درهم‌ریختگی برای ارزیابی موفقیت مدل را ساختند و راه‌های ممکن جدید برای شناسایی وابستگی نژاد حیوانات را معرفی کردند. آن‌ها ابتدا با استفاده از روش‌های مرسوم، مانند نرم‌افزار R و PopGen پارامترهای ژنتیکی جمعیت‌ها، از قبیل پارامترهای تنوع ژنتیکی و پارامترهای همخونی را محاسبه و برآورد کردند. پس از آن خوشه‌بندی جمعیت‌ها با برنامه‌های STRUCTURE و Geneland با به‌کارگیری الگوریتم زنجیره مارکوف مونت کارلو[۱۹] ((MCMC) برای آماره‌های بیزی را انجام دادند. برای ساخت درختان فیلوژنتیک مستقیم، از روش‌های UPGMA و Neighbor-joining استفاده کردند. در بیشتر موارد، تجزیه و تحلیل داده‌ها را با استفاده از بسته‌ها و کتابخانه‌های موجود برای R یا Python انجام دادند. آن‌ها مدل CatBoostClassifier از کتابخانه CatBoost از Yandex را به عنوان مدل آموزش دیده انتخاب کردند. تابع زیان یا هزینه[۲۰] آن‌ها MultiClass بود. (https://catboost.ai/en/docs/concepts/loss-functionsmulticlassification#usage-information) نسبت مجموعه‌های آموزشی و اعتبارسنجی همراه با جابجایی ردیف و هم‌ترازی کلاس (طبقه‌بندی) ۸۰ به ۲۰ بود. آن‌ها گزارش کردند که یکی از مشکلاتی که در فرآیند به کارگیری هوش مصنوعی در علم به وجود می آید، تفسیر نتایج آموزش مدل است. مشکلات به این دلیل به وجود می‌آیند که مدل‌های هوش مصنوعی اغلب بر اساس الگوریتم‌های پیچیده و مقادیر زیادی داده عمل

[18] standard deviation reduction
[19] markov chain monte carlo
[20] loss function

می‌کنند که پردازش آنها برای انسان دشوار است. یادگیری عمیق و یادگیری ماشین رویکردهای متفاوتی را برای ایجاد مدل‌های هوش مصنوعی ارائه می‌کنند (Pour Hamidi et al. 2017). یادگیری ماشینی اغلب از مدل‌های ساده مانند درخت تصمیم یا مدل‌های خطی استفاده می‌کند که تفسیر آن‌ها نسبت به مدل‌های یادگیری عمیق آسان‌تر است. آن‌ها معمولاً پارامترهای کمتری دارند و از ریاضیات ساده‌تری استفاده می‌کنند که درک آن‌ها را برای انسان آسان‌تر می‌کند. مدل‌های یادگیری عمیق مانند شبکه‌های عصبی می‌توانند پیچیده‌تر و دشوارتر برای تفسیر باشند، زیرا لایه‌های پنهان و نورون‌های زیادی در هر لایه دارند. آن‌ها می‌توانند توابع پیچیده را تقریب بزنند، اما این می‌تواند درک نحوه تصمیم گیری آن‌ها را دشوار کند (Ghotbaldini et al. 2019). بنابراین، تفسیرپذیری یک عامل مهم در هنگام انتخاب بین یادگیری عمیق و یادگیری ماشین است. اگر مدل قابل تفسیر بیشتری مورد نیاز است، آن‌گاه مدل یادگیری ماشینی ممکن است انتخاب بهتری باشد. اگر به یک مدل قدرتمندتر نیاز دارید که بتواند وابستگی‌های پیچیده را تقریبی کند، ممکن است بخواهید به یادگیری عمیق روی بیاورید. از نظر ژنتیک و اصلاح نژاد حیوانات و همچنین بسیاری از شاخه های دیگر علم، مشکل انباشت و ایجاد حجم زیادی از داده‌ها وجود دارد. در زمینه دامپروری، عدم وجود مقادیر کافی داده فنوتیپی (داده‌های زئوتکنیکی و دامپزشکی)، داده ژنتیکی (انواع نشانگرها، ژن‌ها، ژنوم‌های توالی یابی شده)، داده عکاسی و داده ویدئویی برای ایجاد پایگاه‌های داده قابل استفاده، که بتوانند با استفاده از روش‌های مختلف هوش مصنوعی مطالعه شوند قابل ذکر است. توجه داشته باشید که قابلیت اطمینان داده‌های به دست آمده به طور مستقیم بر کیفیت مدل‌های آموزش دیده تاثیر می‌گذارد. در آینده نزدیک، روش‌های یادگیری ماشین در ژنتیک و انتخاب، مبنایی برای حل طیف گسترده‌ای از مسائل علمی و عملی مانند ارزیابی ارزش‌های اصلاح نژاد و مخزن ژنی حیوانات و سازگاری، زیست‌پذیری، نوع روان‌شناختی، پتانسیل ژنتیکی؛ پیش بینی استفاده از آنها؛ ایجاد شرایط بهینه نگهداری و تغذیه؛ انتخاب؛ گزینه‌های تلاقی بین و درون نژادی و ایجاد نژادهای جدید یا دستاوردهای انتخاب و... خواهد شد. در پژوهش آن‌ها، استفاده از پایگاه‌های اطلاعاتی ریزماهواره‌ای بر روی دو نوع حیوانات اهلی و مدل‌های یادگیری ماشینی چند کلاسه، شناسایی دقیق نژاد حیوانات، یعنی نژاد اسب و تمایز بین اشکال اهلی و وحشی گوزن شمالی امکان‌پذیر شد. چشم انداز استفاده از روش‌های یادگیری ماشین در انتخاب سنتی، ژنومی، وابسته به نشانگر و اپی ژنتیک، ارزیابی ارزش ژنتیکی، سلامت حیوانات و سازگاری آنها با شرایط مختلف کشاورزی بسیار زیاد است. جستجو برای ژن‌های جدید؛ تجزیه و تحلیل تعامل آن‌ها؛ و غیره به سطح جدیدی می‌رسند، جایی که به سختی می‌توان توانایی‌های هوش مصنوعی را دست بالا گرفت.

بگینگ

بگینگ[۲۱] یک نوع یادگیری گروهی[۲۲] است. بگینگ، که به آن تجمع بوت استرپ نیز می‌گویند، یک روش مجموعه‌ای است که برای مونتاژ چندین نسخه از یک پیش بینی کننده برای به دست آوردن یک پیش بینی کننده قوی انباشته استفاده می شود

[21] bagging
[22] ensemble learning

(Breiman, 1996). با توجه به یک مجموعه آموزشی برچسب‌گذاری شده $(X_1, Y_1) \cdots (X_n, Y_n)$، الگوریتم بگینگ یک تکرار بوت استرپ $(X_1^*, Y_1^*) \ldots (X_n^*, Y_n^*)$، با انتخاب تصادفی نمونه n بار با جایگزینی از مجموعه داده اصلی، و سپس استفاده از آن‌ها به عنوان مجموعه‌های یادگیری جدید برای مدل CART. مدل نهایی با تکرار این مراحل M بار در طول فرآیند یادگیری به دست می‌آید. هنگام پیش‌بینی یک نتیجه عددی، الگوریتم تجمیع میانگین نتیجه همه پیش‌بینی‌کننده‌ها را محاسبه می‌کند. اگر متغیر هدف یک برچسب کلاس[۲۳] باشد، پیش‌بینی‌کننده بسته‌بندی به عنوان اکثریت رای بر روی مدل‌های M تعریف می‌شود (Bühlmann 2012). الگوریتم‌های بگینگ بهتر از مدل‌های CART ساده عمل می‌کنند و دستاوردهای قابل توجهی را در دقت و بهینه‌سازی قابل توجهی برای یادگیرندگان ضعیفی که رفتار ناپایدار از خود بروز می‌دهند، نشان می‌دهند. با این حال، الگوریتم‌های بگینگ به تغییرات در مجموعه‌های آموزشی حساس هستند و می‌توانند کمی عملکرد رویه‌های پایدار را کاهش دهند (Freund and Schapire 1996).

جنگل تصادفی

جنگل تصادفی نوع دیگر یادگیری گروهی[۲۴] است و شامل ترکیبی از پیش‌بینی‌کننده‌های درختی است که به‌عنوان یک مجموعه عمل می‌کند. این درختان تصمیم توسط یک الگوریتم درخت‌سازی تصادفی تولید می‌شوند. الگوریتم چندین درخت را با استفاده از نمونه‌های تصادفی مختلف با همان اندازه مجموعه آموزشی اصلی با گنجاندن موارد خاص بیش از یک بار می‌سازد. علاوه بر این، در هر گره از درخت‌های تصمیم، تقسیم یک زیر مجموعه تصادفی کوچک از ویژگی‌ها را در نظر می‌گیرد. در نتیجه، پیش بینی این درختان می تواند متفاوت باشد. سپس مقدار هدف بر اساس رای اکثریت در مورد پیش‌بینی درختان به کلاس خاصی اختصاص داده می‌شود (Kingsford and Salzberg 2008). از جنگل‌های تصادفی نیز می‌توان برای رگرسیون استفاده کرد که در این صورت مقدار تخمینی متغیر خروجی، میانگین پیش بینی درختان جنگل است.

مثال کاربردی: در پژوهشی .Li et al (2024) مناسب بودن و قابلیت اطمینان روش‌های مختلف یادگیری ماشین برای توسعه مدل‌های پیش‌بینی ژنومی برای صفات اقتصادی مختلف در جوجه‌ها را بررسی کردند. همچنین تأثیر مجموعه‌های اسنیپ انتخاب شده با استفاده از روش‌های بیوانفورماتیک متمایز بر نتایج برازش نهایی را ارزیابی نمودند. برای این امر، آن‌ها از یک تراشه illumina 50K SNPبرای تعیین ژنوتیپ ۴۱۹۰ جوجه ماده نژاد رود آیلند رد تخمگذار استفاده کردند. روش‌های یادگیری ماشین و بیوانفورماتیک کلاسیک برای تناسب ژنوتیپ‌هایی با ۱۰ ویژگی اقتصادی در جوجه‌ها را ادغام کردند. اثربخشی روش‌های یادگیری ماشین را با استفاده از ضرایب همبستگی پیرسون و RMSE بین مقادیر فنوتیپی پیش‌بینی‌شده و واقعی ارزیابی کردند و آنها را با

[23] class label
[24] ensemble learning

rrBLUP[25] و BayesA[26] مقایسه نمودند. آن‌ها از جعبه ابزار منبع باز AutoML AutoGluon v.1.0.0 برای ساخت مدل‌های پیش بینی ژنومی بر اساس الگوریتم های یادگیری ماشین استفاده کردند. مدل‌های یادگیری ماشین مورد استفاده در مطالعه آن‌ها شامل جنگل تصادفی (RF)، درختان مازاد یا جنگل بی‌نهایت تصادفی[۲۷] (ET)، CatBoost (CAT)، K نزدیکترین همسایگان[۲۸] (KNN)، LightGBM (LGB)، شبکه‌های عصبی[۲۹] (NN)، و ساختارهای انباشته‌ای بودند که از این الگوریتم‌ها با برازش خطی به دست می‌آیند و Weighted Ensemble (WE) نامیده می‌شوند. نتایج آن‌ها نشان داد که الگوریتم‌های یادگیری ماشین عملکرد بهتری نسبت به rrBLUP و BayesA در پیش‌بینی وزن بدن و صفات قدرت پوسته تخم مرغ از خود نشان می‌دهند. در مقابل، rrBLUP و BayesA 2 تا ۵۸ درصد دقت پیش‌بینی بالاتری را در پیش‌بینی تعداد تخم‌مرغ نشان دادند. علاوه بر این، ادغام اسنیپ‌های قابل توجه به دست آمده از طریق GWAS در مدل های یادگیری ماشین منجر به افزایش دقت پیش بینی ۰/۱ تا ۲۷ درصد در تقریباً همه صفات شد. این یافته‌ها پتانسیل ترکیب روش‌های بیوانفورماتیک کلاسیک با تکنیک‌های یادگیری ماشین را برای بهبود پیش‌بینی ژنومی در آینده نشان می‌دهد.

بوستینگ

بوستینگ[۳۰] نوع سوم یادگیری گروهی[۳۱] است و یک استراتژی است که برای افزایش دقت مدل‌های پیش بینی استفاده می‌شود. این کار با ادغام چندین مدل ساده، معروف به یادگیرندگان ضعیف، در یک مدل جامع و دقیق‌تر عمل می‌کند. این یادگیرندگان ضعیف، مانند درختان تصمیم‌گیری پایه، به تنهایی قدرت پیش بینی بالایی ندارند. با این حال، هنگامی که بسیاری از آنها با استفاده از یک الگوریتم بوستینگ ترکیب می‌شوند، دقت جمعی آنها به طور قابل توجهی بهبود می‌یابد. یکی از پرکاربردترین الگوریتم‌های بوستینگ کاربردی و عملی Adaboost است. روند یادگیری این الگوریتم با مثال‌های آموزشی با برچسب m شروع می شود:

$$S = ((x_1, y_1) \cdots (x_m, ym))$$

که در آن xi متعلق به فضای X است و به عنوان بردار مقادیر ورودی نشان داده می‌شود، و $yi \epsilon Y$ خروجی برچسب‌دار مرتبط با xi است. الگوریتم بوستینگ به طور مکرر در یک سری دور اجرا می‌شود $t = 1, \ldots, T$ و هر یادگیرنده ضعیفی که به توزیع Dt داده شده است، که به توزیع وزن‌های اختصاص داده شده به مثال‌های مجموعه آموزشی S در هر تکرار اشاره دارد، یک فرضیه ضعیف $ht: X \rightarrow Y$ پیدا می‌کند. هدف کلی الگوریتم یادگیری ضعیف، یافتن فرضیه‌ای به نام فرضیه ضعیف است که خطای

[25] ridge regression BLUP
[26] root-mean-square error
[27] extra tree
[28] K-nearest neighbors
[29] neural networks
[30] boosting
[31] ensemble learning

وزنی t مرتبط با Dt را به حداقل می‌رساند. نتیجه نهایی الگوریتم بوستینگ، ترکیبی از تمام فرضیه‌های ضعیف است، که در آن به هر یک با توجه به اهمیتش وزن (αt) اختصاص داده می‌شود. هر چه یک فرضیه ضعیف دقیق تر باشد، وزن آن بیشتر است. این ترکیب نهایی نوعی «رای اکثریت» از تمام فرضیه‌های ضعیف است و بسیار دقیق‌تر از هر یک از یادگیرندگان ضعیف است. از نظر ریاضی، فرضیه نهایی H به عنوان رای اکثریت وزنی فرضیه‌های ضعیف نشان داده می‌شود، که در آن هر فرضیه ht در وزن αt ضرب می‌شود (Freund and Schapire 1996). بوستینگ در کاهش تنوع تصادفی (واریانس) و خطای سیستماتیک (سوگیری) در پیش بینی‌ها موثر است. همچنین دارای یک ویژگی منحصر به فرد است که در آن بیشتر بر روی نمونه‌های چالش برانگیزتر بر اساس عملکرد یادگیرندگان ضعیف قبلی تمرکز می‌کند. این امر باعث می‌شود الگوریتم‌های بوستینگ بهتر از روش‌های دیگر مانند بگینگ عملکرد بهتری داشته باشند و حساسیت کمتری نسبت به تغییرات داده‌های آموزشی داشته باشند.

مثال کاربردی: در پژوهشی Fadul-Pacheco et al. (2021) روش‌های مختلف طبقه‌بندی یادگیری ماشین، از جمله الگوریتم‌های بگینگ را برای شناسایی گاوهای مثبت به ورم پستان بالینی (CM) در طول اولین شیردهی آنها (اولین شیردهی) و برای پیش‌بینی روزانه شروع ورم پستان بالینی (مداوم) مورد ارزیابی قرار دادند. آن‌ها از داده‌های یکپارچه از منابع مختلف داده از طریق پروژه Dairy Brain در دانشگاه WisconsineMadison استفاده کردند (Cabrera et al. 2020). گاوهایی که برای ورم پستان بالینی مثبت تایید شدند، آن دسته از افرادی بودند که از نظر بالینی بر اساس سوابق بهداشتی مزرعه از جمله نتایج کشت، گرم مثبت، گرم منفی یا سایر عوامل بیماری زا تشخیص داده شدند. در این پژوهش گاوهای بدون سوابق کشت یا آن‌ها که درمان بیشتری دریافت نکردند برای ورم پستان بالینی منفی در نظر گرفته شدند. سایر مطالعات با ورم پستان بالینی نیز کشت‌های منفی را حذف کردند (Ericsson Unnerstad et al. 2009). از جمله نتایج منفی کشت، عدم اطمینان در مورد وضعیت نهایی ورم پستان بالینی، می‌تواند منجر به افزایش مثبت کاذب شود که برای دقت پیش‌بینی مضر است. این پژوهشگران الگوریتم جنگل تصادفی را با روش بوت استرپ بگینگ (انباشتگی) آزمایش کردند. این روش برای کاهش واریانس یک الگوریتم استفاده می‌شود. این یک مشکل رایج در جنگل تصادفی است، زیرا آن‌ها به تغییرات در داده‌های آموزشی بسیار حساس هستند. بگینگ با درخت‌های تصمیم‌گیری مانند جنگل تصادفی، واریانس و مشکل بیش از حد برازش داده‌ها را کاهش می‌دهد (Brownlee 2016). از آنجایی که بالاترین هزینه در پژوهش آن‌ها برای منفی کاذب یا حیوانات مبتلا به ورم پستان بالینی بود اما به‌عنوان منفی شناخته می‌شدند، گاوهای منفی کاذب را با جزئیات دنبال کردند و نسبت recall را هم محاسبه کردند.

$$Recall = \frac{true\ cases\ positive}{true\ cases\ positive + false\ negatives}$$

recall برای حیوانات مثبت ۶۷ درصد بود. در مقابل، هنگام ترکیب روش بگینگ با random forest-wrapper، تعداد کلی نمونه‌های به درستی طبقه‌بندی شده ۶۶ درصد و سطح وزنی AUC-ROC برابر ۷۴ درصد و recall برای حیوانات مثبت

۶۷ درصد بود. برای بدست آوردن جنگل تصادفی، شاخص نمونه‌های طبقه‌بندی صحیح ۶۰ درصد و ناحیه وزنی AUC-ROC برابر ۶۷ درصد بود recall برای حیوانات مثبت ۵۷ درصد بود که نتایج به‌دست‌آمده با random forest-wrapper را کمتر انجام داد. آن‌ها گزارش دادند، اگرچه نتایج الگوریتم‌های Naïve Bayes و جنگل تصادفی[۳۲] برای پیش‌بینی شروع ورم پستان بالینی در طول شیردهی اول، زمانی که فقط داده‌های مربوط به ویژگی‌های ژنتیکی را شامل می‌شد، آنطور که انتظار می‌رفت خوب نبود، نتایج با گنجاندن داده‌های ژنتیکی، سلامتی و تولیدی یکپارچه دلگرم‌کننده بود. نتایج الگوریتم جنگل تصادفی، به ویژه با ویژگی‌های انتخاب شده توسط روش Wrapper، امیدوارکننده بود. برای پیش‌بینی مداوم روزانه ورم پستان بالینی، نتایج نشان داد که الگوریتم جنگل تصادفی بهتر از Extreme Gradient Boosting پیش‌بینی می‌کند که به خوبی موارد ورم پستان بالینی را قبل از شروع پیش‌بینی می‌کند. به همین ترتیب، ادغام سایر جریان‌های داده به‌عنوان حسگر، می‌تواند به بهبود قدرت پیش‌بینی و به حداکثر رساندن استفاده از رکوردهای جمع‌آوری‌شده در مزرعه برای تولید الگوریتم‌های مبتنی بر مزرعه با توجه به نیازهای کشاورز و شرایط مدیریت کمک کند. علاوه بر این، داشتن دو الگوریتم مختلف یادگیری ماشین ممکن است در تصمیم گیری کوتاه مدت و میان مدت کمک کند. یکی از الگوریتم‌ها می‌تواند برای شناسایی گاوهایی که به طور کلی در معرض خطر ابتلا به ورم پستان بالینی در اولین شیردهی هستند، استفاده شود، در حالی که الگوریتم دیگر می‌تواند برای پیش‌بینی ورم پستان بالینی روزانه استفاده شود، که می‌تواند گاو را در معرض خطر در هر دوشش نشان دهد. شناسایی گاوهای با خطر بالاتر ورم پستان بالینی می‌تواند برای کشاورز برای انجام اقدامات به موقع که می‌تواند از اثرات منفی ورم پستان بالینی پیشگیری کرده و آن‌ها را بهبود بخشد، سودمند باشد.

فضای هیلبرت با هسته بازآفرین

فضای هیلبرت با هسته بازآفرین[۳۳] (RKHS)، یکی از الگوریتم‌های مبتنی بر هسته[۳۴] است و یک مدل رگرسیون نیمه پارامتریک است که برای اولین بار بر روی ژنوتیپ‌های نشانگر توسط Gianola et al. (2011) اعمال شد. این روش پتانسیل محاسباتی زیادی را نشان داده است، به خصوص زمانی که $p >> n$ است. RKHS یک فضای هیلبرت (H) از توابع است که در آن هر تابع را می توان به عنوان نقطه‌ای در فضای اقلیدسی در نظر گرفت، و حدی و خطی فرض می‌شود. به عبارت دیگر، اگر دو تابع f و g دارای هنجارهای نزدیک $\|f(x) - g(x) \to 0\|$ باشند، آنها نیز مقادیر نزدیک $|f(x) - g(x) \to 0|$ دارند. وظیفه یادگیری RKHS را می‌توان به صورت زیر توصیف کرد:

[32] random forest
[33] reproducing kernel hilbert spaces
[34] Kernel-based algorithms

فرض کنید x_i یک بردار ژنوتیپ‌های نشانگر (ورودی)، y_i یک بردار مقادیر ژنتیکی (خروجی) و $g(x)$ یک تابع ناشناخته از اثرات ژنتیکی باشد. برای استنباط g، RKHS با تعریف فضایی از توابع پیش می‌رود که اگر تابع تلفات زیر را به حداقل برساند، عنصر $\hat{g}$ از آن انتخاب می‌شود:

$$l(g|\lambda) = \|y - g\|^2 + \lambda\|g\|_H^2$$

که در آن λ یک پارامتر منظم‌سازی است که مبادلات بین خوبی تناسب و پیچیدگی مدل را کنترل می‌کند، H نشان‌دهنده فضای هیلبرت است، و $\|g\|_H^2$ مربع هنجار g در H است. مربع هنجار پیچیدگی مدل را اندازه‌گیری می‌کند. با توجه به نظریه Manton and Amblard (2014)، RKHS می‌تواند برای حل سه نوع مسئله استفاده شود:

الف) زمانی که مشکل روی یک فضای فرعی که اتفاقاً RKHS است، تعریف می‌شود. این نشان می‌دهد که نگاشت فضای مسئله در فضایی با ابعاد بالاتر، مسئله را آسان‌تر می‌کند. انتخاب ژنومی چالشی با ابعاد بالا ایجاد می‌کند زیرا تعداد ژنوتیپ ها (p) معمولاً از تعداد افراد (n) بیشتر می‌شود. با استفاده از چارچوب RKHS، کاهش این ابعاد و تسهیل حل چنین مشکلاتی امکان پذیر می‌شود. معرفی یک هسته گاوسی امکان تبدیل داده‌های ژنوتیپی به یک نمایش RKHS مناسب را فراهم می‌کند، که به موجب آن مدل‌های رگرسیون خطی بعدی می‌توانند به طور موثر برای پیش‌بینی مقادیر ژنتیکی در این فضای کاهش بعدی استفاده شوند.

ب) هنگامی که یک مسئله دارای عملکرد نیمه قطعی مثبت است. در زمینه انتخاب ژنومی، یک جزء مهم ماتریس رابطه ژنتیکی است (که به عنوان ماتریس خویشاوندی نیز شناخته می‌شود)، که شباهت ژنتیکی بین افراد را به صورت کمی نشان می‌دهد. این تابع هدف مهمی در تصحیح عوامل مخدوش کننده مانند ساختار جمعیت و وابستگی خانوادگی در مطالعات همبستگی دارد. استفاده از فضای هیلبرت با هسته بازآفرین یکی از راه حل‌های مشکلی است که ژنوتیپ‌های با ابعاد بالا ارائه می‌کنند. با استفاده از این رویکرد، می‌توانیم از ترفند هسته برای مدیریت مؤثر و قابل کنترل‌تر کردن این مشکل پیچیده استفاده کنیم.

ج) زمانی که می‌توان نقاط داده را در یک RKHS جاسازی کرد که تابع هسته ویژگی‌های تابع فاصله را با توجه به تمام نقاط داده و تابعی که فاصله بین آنها را تعیین می‌کند، ثبت می‌کند (Nayeri et al. 2019). یکی از وظایف رایج در انتخاب ژنومی، گروه بندی افراد بر اساس ژنوتیپ آن‌هاست. این معمولاً برای اهدافی مانند شناسایی زیرجمعیت‌ها یا بررسی ساختار جمعیت انجام می‌شود. برای دستیابی به این هدف، ژنوتیپ‌ها را می‌توان با استفاده از یک تابع هسته مناسب، مانند یک هسته گاوسی یا خطی، در یک فضای هیلبرت هسته قابل تکرار قرار داد. با انجام این کار، می‌توانیم شباهت ژنتیکی افراد را به تصویر بکشیم. الگوریتم خوشه‌بندی در این RKHS عمل می‌کند و هدف آن یافتن خوشه‌هایی است که به خوبی در RKHS از هم جدا شده‌اند، حتی اگر ممکن است در فضای ژنوتیپ اصلی به خوبی از هم جدا نشوند.

ماشین‌های بردار پشتیبان

ماشین‌های بردار پشتیبان[35] (SVM) یکی دیگر از الگوریتم‌های مبتنی بر هسته[36] است که یک الگوریتم ناپارامتریک است که توسط Cortes and Vapnik (1995) پیشنهاد شده است. برای اولین بار برای مسائل طبقه بندی دو گروهی معرفی شد. با این حال، امروزه به طور گسترده‌ای برای رگرسیون و طبقه بندی استفاده می‌شود. هنگام برخورد با خوشه بندی، هدف الگوریتم SVMشناسایی یک ابر صفحه بهینه تعریف شده به عنوان مرزی است که حداکثر کلاس‌ها را از هم جدا می‌کند (Jiang et al. 2020). هنگامی که نقاط داده به صورت خطی قابل تفکیک هستند، الگوریتم SVM یک طبقه بندی خطی را انجام می‌دهد و ابر صفحه بهینه با استفاده از بهینه سازی عددی پیدا می‌شود (Crisci et al. 2012). در غیر این صورت، SVM می‌تواند با استفاده از تابع Kernel یک طبقه بندی غیر خطی انجام دهد. تابع هسته گاوسی برای نگاشت نقاط داده از یک فضای داده به یک فضای ویژگی با ابعاد بالا استفاده می‌شود. در فضای ویژگی، کره‌های کوچک به نظر می‌رسد که تصویر داده‌ها را محصور می‌کنند، این کره‌ها به فضای داده نگاشت می‌شوند و مرزهای خوشه‌ای را تشکیل می‌دهند که نقاط داده همان خوشه را در بر می‌گیرد (Ben-Hur et al. 2001). مرزها باید حاشیه بین آن‌ها و طبقات را به حداکثر برسانند تا خطای طبقه بندی را به حداقل برسانند (Mahesh 2020). هنگامی که الگوریتم SVM برای مسائل رگرسیون اعمال می شود، تابع ضرر باید شامل اندازه‌گیری فاصله باشد. توابع زیان احتمالی عبارتند از تابع افت درجه دوم، لاپلاسی، هوبر و تابع ضرر غیر حساس (Gunn 1998). الگوریتم‌های SVM به دلیل انعطاف‌پذیری می‌توانند پیش‌بینی‌های بسیار دقیقی داشته باشند. با این حال، آن‌ها به عنوان یک جعبه سیاه توصیف می‌شوند، زیرا هیچ معیاری برای چگونگی بهینه سازی ابرصفحه توسط پیش بینی کننده‌ها ارائه نشده است، که تفسیر پیش بینی‌ها را سخت می‌کند.

مثال کاربردی: در پژوهشی Wang et al. (2022) از یادگیری ماشین برای بهبود دقت پیش‌بینی ژنومی صفات تولید مثلی در خوک استفاده کردند. اهداف آن‌ها ارزیابی عملکرد روش‌های یادگیری ماشین در پیش‌بینی ژنومی در مقایسه با روش‌های رایج موجود GBLUP[37]، ssGBLUP[38] و BayesHE و ارزیابی کارایی روش‌های مختلف یادگیری ماشین برای کشف روش یادگیری ماشین ایده‌آل برای پیش‌بینی ژنومی بود. در این مطالعه، ۲۵۶۶ خوک یورکشایر چینی با سوابق صفات تولید مثلی با پانل‌های GenoBaits Porcine SNP 50 K و PorcineSNP50 تعیین ژنوتیپ شدند. چهار روش یادگیری ماشین شامل رگرسیون بردار پشتیبان (SVR)، رگرسیون ریج کرنل (KRR)، جنگل تصادفی (RF) و Adaboost.R2 اجرا شد. از طریق ۲۰ تکرار اعتبارسنجی متقابل (CV) پنج برابری و یک پیش‌بینی برای افراد جوان‌تر، کاربرد روش‌های یادگیری ماشین در پیش‌بینی ژنومی مورد بررسی قرار گرفت.

مدل GBLUP به صورت زیر بود:

[35] support vector machines
[36] kernel-based algorithms
[37] genomic best linear unbiased prediction
[38] single-step genomic best linear unbiased prediction

$$y_c = 1\mu + Zg + e$$

که در آن y_c ناقل فنوتیپ‌های اصلاح شده افراد تعیین ژنوتیپ شده است، μ میانگین کلی، ۱ بردار 1 s، g بردار مقادیر اصلاح ژنومی، e بردار خطاهای تصادفی، و Z یک ماتریس بروز است که رکوردها را به g اختصاص می‌دهد. توزیع اثرات تصادفی عبارت بودند از $g \sim N(0, G\,\sigma_g^2)$ و $e \sim N(0, I\,\sigma_e^2)$، که در آن G ماتریس همبستگی یا رابطه ژنومی (ماتریس G)، و σ_g^2 و σ_e^2 به ترتیب واریانس ژنتیکی افزایشی و واریانس باقی مانده هستند.

مدل ssGBLUP: ssGBLUP همان بیان GBLUP را داشت، با این تفاوت که از y_c افراد تعیین ژنوتیپ شده و تعیین ژنوتیپ نشده با ترکیب ماتریس G و ماتریس A استفاده می‌کرد. فرض بر این بود که g از توزیع نرمال $N(0, H\,\sigma_g^2)$ پیروی می‌کند. معکوس ماتریس H به صورت زیر بود:

$$H^{-1} = \begin{bmatrix} G_w^{-1} - A_{22}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + A^{-1}$$

برای جلوگیری از مشکل معکوس نشدن ماتریس منفرد $G_w = (1\text{-}w)\,G_a + wA_{22}$ و w برابر ۰/۰۵ بود (Forni et al. 2011).

مدل BayesHE: در این پژوهش BayesHE توسط Shi et al. (2021) توسعه یافت. این کار برای افزایش انعطاف‌پذیری و انطباق‌پذیری مدل بیزی، بر اساس اولویت‌های جهانی-محلی بود. در این مطالعه، اولین شکل BayesHE (BayesHE1) را مورد استفاده قرار دادند، و زنجیره مارکوف مونت کارلو (MCMC) برای ۵۰۰۰۰ چرخه اجرا شد که ۲۰۰۰۰ چرخه اول به عنوان سوزاندن آزمایشی دور انداخته شد و هر ۵۰ نمونه از آن ۳۰۰۰۰ باقی‌مانده برای استنباط آمارهای قبلی ذخیره شد. اسکریپت‌های داخلی نوشته شده در Fortran 95 برای تجزیه و تحلیل BayesHE و روش DMUAI پیاده سازی شده در نرم افزار DMU (Madsen et al. 2014) برای تجزیه و تحلیل GBLUP و ssGBLUP استفاده شد.

رگرسیون بردار پشتیبان: ماشین بردار پشتیبان (SVM) بر اساس تئوری یادگیری آماری است. SVR کاربرد SVM در رگرسیون برای مقابله با پاسخ‌های کمی بود که از یک تابع هسته خطی یا غیرخطی برای ترسیم فضای ورودی (مجموعه داده نشانگر) به فضای ویژگی ابعاد بالاتر استفاده می‌کرد (Müller and Guido 2017)، و مدل‌سازی و پیش‌بینی را روی فضای ویژگی انجام می‌داد. به عبارت دیگر، می‌توانیم یک مدل خطی در فضای ویژگی برای مقابله با مشکلات رگرسیون بسازیم. فرمول مدل SVR را می‌توان به صورت زیر بیان کرد:

$$f(x) = \beta_0 + h(x)^T \beta$$

که در آن $h(x)^T\beta$ تابع هسته، β بردار وزن‌ها و β0 بایاس (اریب) است. به طور کلی، SVR فرموله شده شده با به حداقل رساندن تابع ضرر محدود شده زیر به دست می‌آید:

$$\min_{\beta_0,\beta} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n} V(y_i - f(x_i))$$

که در آن

$$V_\varepsilon(r) = \begin{cases} 0, if\,|r| < \varepsilon \\ |r| - \varepsilon, otherwise \end{cases}$$

که در آن $V_\varepsilon(r)$ تلفات حساس به ε و C ("پارامتر هزینه") ثابت منظم سازی است که مبادله بین خطای پیش بینی و پیچیدگی مدل را کنترل می‌کند. y یک پاسخ کمی است و ||·|| هنجار در فضای هیلبرت است. پس از بهینه سازی، شکل نهایی SVR را می توان به صورت زیر نوشت:

$$f(x) = \sum_{i=1}^{m} (\hat{a}_i - a_i) k(x, x_i)$$

که در آن $k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ تابع هسته است. در این پژوهش از جستجوی شبکه‌ای برای یافتن بهترین تابع هسته و فراپارامترهای بهینه C و گاما استفاده کردند. یک استراتژی اعتبارسنجی متقابل پنج برابری ((5-fold CV) برای تنظیم فراپارامترها هنگام انجام جستجوی شبکه‌ای انجام دادند.

رگرسیون ریج کرنل (KRR): رگرسیون ریج کرنل (KRR) یک روش رگرسیون غیرخطی است که می‌تواند به طور موثر ساختار غیرخطی داده‌ها را کشف کند KRR از یک تابع هسته غیرخطی برای نگاشت داده‌ها به فضای هسته با ابعاد بالاتر استفاده می‌کند و سپس یک مدل رگرسیون بریج ایجاد می‌کند تا داده‌ها را به صورت خطی در این فضای هسته جدا کند. تابع خطی در فضای هسته با توجه به میانگین مجذور افت خطای منظم‌سازی برآمدگی انتخاب شد (Exterkate et al. 2016). مدل نهایی پیش‌بینی KRR را می‌توان به صورت زیر نوشت:

$$y(x_i) = k'(K + \lambda I)^{-1}\hat{y}$$

که در آن λ ثابت منظم‌سازی است و K ماتریس Gram با ورودی‌های $K_{ij} = K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)^T$. بنابراین، برای nنمونه آموزشی، ماتریس kernel بدست آمده به صورت زیر است:

Agricultural Biotechnology Journal; Print ISSN: 2228-6705, Electronic ISSN: 2228-6500

$$K = \begin{bmatrix} K(x_1,x_1) & K(x_1,x_2) & \cdots & K(x_1,x_n) \\ K(x_2,x_1) & K(x_2,x_2) & \cdots & K(x_2,x_n) \\ \vdots & \vdots & \vdots & \vdots \\ K(x_n,x_1) & K(x_n,x_2) & \cdots & K(x_n,x_n) \end{bmatrix}_{n*n}$$

I ماتریس همانی (یکه) است، $k^{'} = K(x_i, x_j)$ با $j = 1,2,3, \ldots,n,$ ، n تعداد نمونه‌های آموزشی و x_i نمونه آزمایشی است. شکل گسترش یافته آن به صورت زیر است:

$$k = \begin{bmatrix} K(x_i,x_1) \\ K(x_i,x_2) \\ \vdots \\ K(x_i,x_n) \end{bmatrix}$$

از جستجوی شبکه برای یافتن مناسب‌ترین تابع هسته و λ در این مطالعه استفاده کردند و یک استراتژی اعتبارسنجی متقابل پنج برابری برای تنظیم فراپارامترها استفاده نمودند.

جنگل تصادفی: جنگل تصادفی (RF) یک روش یادگیری ماشین است که از رای‌گیری یا میانگین درختان تصمیم‌گیری چندگانه برای تعیین طبقه بندی یا مقادیر پیش بینی شده نمونه‌های جدید استفاده می‌کند (Breiman 2001). در این پژوهش جنگل تصادفی اساساً مجموعه‌ای از درختان تصمیم بود و هر درخت تصمیم کمی با درختان دیگر متفاوت بود. این جنگل تصادفی با میانگین‌گیری نتایج پیش‌بینی بسیاری از درختان تصمیم‌گیری، خطر بیش از حد برازش را کاهش داد (González-Camacho et al. 2018). رگرسیون جنگل تصادفی را می توان به شکل زیر نوشت:

$$y = \frac{1}{M}\sum_{m=1}^{M} t_m(\psi_m(y:X))$$

که در آن y مقدار پیش‌بینی‌شده رگرسیون تصادفی جنگل است، $t_m(\psi_m(y:X))$ یک درخت رگرسیون فردی است و M تعداد درخت‌های تصمیم در جنگل است. در پژوهش آن‌ها پیش‌بینی با انتقال متغیرهای پیش‌بینی‌کننده در فلوچارت هر درخت به دست آمد و مقدار تخمینی مربوطه در گره پایانی به عنوان مقدار پیش‌بینی‌شده استفاده شد. در نهایت، پیش‌بینی‌های هر درخت در RFبرای محاسبه پیش‌بینی نهایی داده‌های مشاهده نشده میانگین‌گیری شد. جستجوی شبکه‌ای برای یافتن مناسب ترین هایپرپارامتر M و حداکثر عمق درخت و اعتبارسنجی متقابل پنج برابری برای تنظیم هایپرپارامترها انجام شد.

Adaboost.R2: Adaboost.R2 یک اصلاح موردی از Adaboost است. R و یک فرمت Adaboost.M2 ایجاد شده برای مقابله با مشکلات رگرسیون، که به طور مکرر از درخت رگرسیون به عنوان یک یادگیرنده ضعیف استفاده می‌کند و به

دنبال آن وزن نمونه‌های پیش‌بینی نادرست را افزایش می‌دهد و وزن نمونه‌های پیش‌بینی‌شده درست را کاهش می‌دهد. با ادغام چند یادگیرنده ضعیف یک «کمیته» ایجاد می‌کند و تأثیر پیش‌بینی آن را بهتر از یادگیرندگان ضعیف می‌کند (Shrestha and Solomatine 2006). مدل رگرسیون Adaboost.R2 را می‌توان به صورت زیر نوشت:

$$y = \inf\left[y \in Y : \sum_{t:f_t(x)\le y} \log\frac{1}{\varepsilon_t} \ge \frac{1}{2}\sum_t \log\frac{1}{\varepsilon_t}\right]$$

که در آن y مقدار پیش‌بینی‌شده، $f_t(x)$ مقدار پیش‌بینی‌شده t امین یادگیرنده ضعیف، ε_t میزان خطای $f_t(x)$ و $\varepsilon_t=$ $\overline{L}_t = \overline{L}_t \div (1-\overline{L}_t)$، $\overline{L}_t$ میانگین ضرر و $\overline{L}_t = \sum_{i=1}^{m} L_t(i)\ D_t(i)$ است، $L_t(i)$ خطای بین مقدار واقعی مشاهده و مقدار پیش بینی شده-i امین فرد پیش بینی شده است و $D_t(i)$ توزیع وزنی $f_t(x)$ است. پس از این که $f_t(x)$ آموزش داده شد توزیع وزنی $D_t(i)$ تبدیل به $D_{t+1}(i)$ می‌شود.

$$D_{t+1}(i) = \frac{D_t(i)\beta_t^{(1-L_t(i))}}{Z_t}$$

که در آن Z_t یک عامل عادی سازی است که به گونه‌ای انتخاب شده است که $D_{t+1}(i)$ یک توزیع باشد. در این مطالعه پژوهشگران از SVR و KRR به عنوان یادگیرندگان ضعیف Adaboost.R2 استفاده کردند. برای این کار چهار روش یادگیری ماشین، وکتورهای ژنوتیپ‌ها (کد شده به صورت ۰، ۱، ۲) متغیرهای مستقل ورودی بودند، فنوتیپ‌های تصحیح شده y_c به عنوان متغیر پاسخ و بسته Sklearn برای Python نسخه (V0.22) برای پیش‌بینی ژنومیک استفاده شدند. آن‌ها ترکیب فراپارامتر بهینه را از شبکه‌ای از مقادیر با ترکیب‌های فراپارامترهای مختلف جستجو کردند و ترکیبی در شبکه با بالاترین همبستگی پیرسون به عنوان فراپارامتر بهینه در هر برابر (جستجوی شبکه‌ای) انتخاب نمودند.

نتایج آن‌ها نشان داد که در اعتبارسنجی متقابل (CV)، در مقایسه با BLUP ژنومی (GBLUP)، GBLUP تک مرحله‌ای (ssGBLUP) و روش BayesHE، روش‌های یادگیری ماشین به طور قابل‌توجهی از این روش‌های مرسوم بهتر عمل کردند. روش‌های یادگیری ماشین دقت پیش‌بینی ژنومی GBLUP، ssGBLUP و BayesHE را به ترتیب ۱۹/۳، ۱۵ و ۲۰/۸ درصد بهبود بخشید. علاوه بر این، روش‌های یادگیری ماشین میانگین مربعات خطای (MSE) و میانگین خطای مطلق (MAE) کوچک‌تری را در همه سناریوها به همراه داشتند. ssGBLUP در مقایسه با GBLUP به طور متوسط ۳/۸ درصد بهبودی در دقت داشت و دقت BayesHE نزدیک به GBLUP بود. در پیش‌بینی ژنومی افراد جوان‌تر، RF و Adaboost.R2_KRR بهتر از GBLUP و BayesHE عمل کردند، در حالی که ssGBLUP در مقایسه با RF و ssGBLUP دقت کمی بالاتر و MSE پایین‌تری نسبت به Adaboost.R2_KRR در پیش‌بینی تعداد کل بچه خوک‌های متولد شده داشت. برای تعداد خوکچه‌های زنده متولد شده، Adaboost.R2_KRR به طور قابل توجهی بهتر از ssGBLUP عمل کرد. در میان روش‌های یادگیری ماشین،

Adaboost.R2_KRR به طور مداوم در این مطالعه عملکرد خوبی داشت. یافته‌های این پژوهشگران همچنین نشان داد که فراپارامترهای بهینه برای روش‌های یادگیری ماشین مفید هستند. آن‌ها نشان دادند که پس از تنظیم فراپارامترها در CV و در پیش‌بینی نتایج ژنومی افراد جوان‌تر، میانگین بهبود به ترتیب ۱۴/۳ درصد و ۲۱/۸ درصد نسبت به افرادی که از فراپارامترهای پیش‌فرض استفاده می‌کردند بود. یافته‌های آن‌ها همچنین نشان داد که تنظیم فراپارامترها برای روش‌های یادگیری ماشین ضروری است و فراپارامترهای بهینه به ویژگی‌های صفات، مجموعه داده‌ها و غیره بستگی دارد.

نزدیکترین همسایگان

مدل نزدیکترین همسایگان یکی از ساده ترین و شهودی ترین الگوریتم‌های یادگیری ماشین است. ایده این رویکرد پیش بینی مقدار متغیر هدف y_i مرتبط با متغیر ورودی x_i بر اساس فاصله بین x_i و سایر نقاط داده است. به طور کلی از فاصله اقلیدسی استفاده می‌شود، اما روش‌های دیگری، مانند فاصله منهتن[۳۹] برای محاسبه این فاصله وجود دارد. در طبقه بندی، y_i به برچسب کلاس اکثر نزدیکترین نقاط داده در فضا اختصاص داده می‌شود. متناوباً، هنگام برخورد با رگرسیون، پیش‌بینی‌کننده میانگین خروجی از نزدیک‌ترین همسایگان است. K نزدیکترین همسایگان[۴۰] (KNN) محبوب‌ترین الگوریتم در این دسته است. این بر اساس همان ایده است که نزدیکترین الگوها به نقطه داده x_i اطلاعات برچسب مفیدی را ارائه می‌دهند. پارامتر ناشناخته K تعیین می‌کند که چه تعداد همسایه در فرآیند یادگیری در نظر گرفته شود. تعداد همسایگان K تأثیر قابل توجهی بر عملکرد الگوریتم دارد. K بهینه آن چیزی است که بین برازش (بایاس کم، اما واریانس زیاد) و عدم تناسب (واریانس کم، اما بایاس زیاد) تعادل برقرار می‌کند. برخی از نویسندگان K را به جذر تعداد مشاهدات در مجموعه آموزشی پیشنهاد می‌کنند (Zhang 2016).

شبکه های عصبی عمیق

یادگیری عمیق خانواده‌ای از روش‌های یادگیری قدرتمند است که قادر به تشخیص الگوهای پیچیده در داده‌های خام است (Vieira et al. 2020). پرسپترون معروف روزنبلات که در دهه ۱۹۵۰ پیشنهاد شد، اولین تلاش برای درک مدلی بود که تقریباً مشابه فرآیندهای ادراکی مغز انسان بود. ساختار شبکه‌های عصبی عمیق[۴۱] (DNN) از لایه‌های روی هم از نورون‌های متصل تشکیل شده است. به عبارت دیگر، مدل DNN شامل تعداد مشخصی لایه است که هر لایه حاوی چندین نورون است. هر نورون از طریق وزنه‌هایی (آرایه‌هایی) به نورون‌های لایه‌های مجاور متصل می‌شود که قدرت و جهت اتصال (تحریکی یا مهاری) را منعکس

[39] Manhattan distance
[40] K-nearest neighbors
[41] deep neural networks

Agricultural Biotechnology Journal; Print ISSN: 2228-6705, Electronic ISSN: 2228-6500

می‌کند. مدل‌های DNN با عمق، اندازه و عرض مشخص می‌شوند. به تعداد لایه‌هایی که در یک DNN وجود دارد، به استثنای لایه ورودی، عمق می‌گویند. تعداد کل نورون‌ها در مدل اندازه نامیده می‌شود. در نهایت، عرض DNN لایه‌ای است که بیشترین تعداد نورون‌ها را در بر می‌گیرد (Montesinos-López et al. 2021). هنگام اجرای DNN، مجموعه‌ای از مشاهدات X از طریق لایه ورودی وارد مدل می‌شود. مشاهدات x_i ورودی و خروجی این لایه هستند. در لایه‌های پنهان DNN، هر نورون از یک لایه معین، از لایه سطح سلسله مراتبی پایین‌تر، مجموع وزنی خروجی نورون‌های خود را دریافت می‌کند و سپس آن را از یک تابع فعال‌سازی عبور می‌دهد تا آن را به عنوان خروجی برای آن نورون هدایت کند. در لایه‌های پنهان، پرکاربردترین توابع فعال‌سازی عبارتند از: واحد خطی اصلاح‌شده، فعال‌سازی مماس هذلولی و تابع سیگموئید. در لایه خروجی، DNN برای انجام یک طبقه‌بندی یا یک رگرسیون بر اساس ماهیت متغیر هدف است. هنگامی که با طبقه بندی سروکار داریم، تعداد نورون‌ها در لایه خروجی با تعداد کلاس‌ها برابر است. علاوه بر این، توابع فعال سازی متفاوتی را می‌توان با توجه به نوع متغیر هدف مورد استفاده قرار داد. Softmax برای متغیرهای طبقه‌بندی، تابع نمایی برای داده‌های شمارش و تابع سیگموئید برای نتایج باینری استفاده می‌شود. در مسائل رگرسیون، لایه خروجی مقادیر تخمینی متغیرهای هدف را نشان می‌دهد و توابع فعال‌سازی خطی اعمال می‌شوند. موفق ترین تابع فعال سازی هنگام برخورد با یک متغیر پیوسته، واحد خطی یکسو شده[۴۲] (ReLU) است (Bircanoğlu and Arıca 2018). تابع فعال‌سازی tanh در DNN برای معرفی غیرخطی بودن در مدل استفاده می‌شود و به مدل اجازه می‌دهد تا از وزن‌های مثبت و منفی یاد بگیرد، زیرا حول محور صفر است (برخلاف تابع سیگموئید). این تابع معمولاً در لایه‌های مخفی استفاده می‌شود.

مانند سایر مدل های یادگیری ماشین، آموزش DNN شامل انتخاب وزن‌های بهینه است که تفاوت بین مقادیر واقعی و تخمینی متغیر هدف را به حداقل می‌رساند. گرادیان کاهشی[۴۳] برای به حداقل رساندن تابع تلفات استفاده می‌شود. این پارامترها باید در طول فرآیند یادگیری به روز شوند. هنگام آموزش برای اولین بار مدل DNN، وزن‌ها به طور تصادفی مقداردهی اولیه می‌شوند. هنگامی که یک مشاهده وارد مدل شد، اطلاعات از طریق شبکه منتشر می شود تا زمانی که مقدار خروجی خاصی را پیش بینی کند. سپس گرادیان‌های تابع از دست دادن با استفاده از یک فراپارامتر به نام نرخ یادگیری η محاسبه می‌شود که نشان می‌دهد مراحل گرادیان کاهشی چقدر باید بزرگ باشد و سپس برای به روز رسانی پارامترهای تابع (وزن‌ها و بایاس‌ها) استفاده می‌شود. پس انتشار[۴۴] روش کارآمد دیگری برای محاسبه گرادیان است. مفهوم این روش بر این اساس استوار است که سهم هر نورون در تابع از دست دادن متناسب با وزن اتصال آن با نورون‌های لایه زیر است. بنابراین، این مشارکت‌ها را می‌توان با شروع از لایه خروجی محاسبه کرد و با استفاده از وزن‌ها و مشتق تابع فعال‌سازی در شبکه منتشر شد. یادگیری عمیق شامل طیف گسترده‌ای از معماری‌ها است.

[42] rectified linear unit
[43] gradient descent
[44] backpropagation

محبوب‌ترین آنها شبکه‌های پیشخور هستند که پرسپترون چندلایه[45] (MLP)، شبکه‌های عصبی بازگشتی[46] (RNN) و شبکه‌های عصبی کانولوشنال[47] (CNN) نیز نامیده می‌شوند.

الف) پرسپترون چندلایه (MLP): پرسپترون چندلایه (MLP) یک شبکه پیشخور لایه‌ای است که در آن تمام لایه‌ها به طور کامل به هم متصل هستند. هر نورون از یک لایه مشخص به نورون‌های لایه مجاور متصل است، اطلاعات در یک جهت واحد جریان می‌یابد. به عبارت دیگر اتصالات درون لایه یا فوق لایه وجود ندارد. MLPها قدرتمند هستند و آموزش آنها ساده است. با این حال، این شبکه‌ها برای مقابله با مجموعه داده‌های مکانی یا زمانی مناسب نیستند و در معرض بیش از حد برازش هستند.

ب) شبکه‌های عصبی بازگشتی (RNN): در شبکه‌های عصبی بازگشتی اطلاعات در هر دو جهت جریان دارد. هر نورون دارای سه نوع اتصال است: اتصالات ورودی از لایه قبلی، اتصالات مداوم به لایه بعدی و اتصالات مکرر بین نورون‌های همان لایه. این ساختار بازگشتی به این شبکه اجازه می‌دهد تا تصوری از حافظه داشته باشد زیرا خروجی یک لایه به ورودی‌های فعلی و قبلی بستگی دارد. RNN اغلب برای مدل‌سازی ساختارهای فضا–زمانی[48] استفاده می‌شود. همچنین در زمینه‌های پردازش زبان طبیعی و تشخیص گفتار استفاده می‌شود.

ج) شبکه‌های عصبی کانولوشنال (CNN): شبکه‌های عصبی کانولوشنال (CNN) برای تطبیق موقعیت‌هایی طراحی شده‌اند که داده‌ها در قالب آرایه‌های متعدد نمایش داده می‌شوند. متغیر ورودی می‌تواند یک بعدی مانند اسنیپ‌ها، دو بعدی مانند تصاویر رنگی، یا سه بعدی برای فیلم ها یا تصاویر حجمی داشته باشد. معماری CNNها از لایه‌های کانولوشن و ادغام و به دنبال آن شبکه‌های عصبی کاملاً متصل تشکیل شده است. هنگام آموزش CNNها، دو نوع لایه اول، یعنی لایه‌های کانولوشن و تلفیقی، استخراج ویژگی را انجام می‌دهند. شبکه عصبی کاملا متصل برای انجام طبقه بندی یا وظیفه رگرسیون است. در لایه کانولوشن، یک عملیات ریاضی برای تولید یک نسخه فیلتر شده از ماتریس‌های اصلی داده‌های ورودی انجام می‌شود. این عملیات کانولوشنی "هسته (کرنل)" یا "فیلتر" نامیده می‌شود. یک تابع فعال‌سازی غیرخطی، معمولاً ReLU، پس از هر پیچیدگی برای تولید خروجی اعمال می‌شود که به صورت نقشه‌های ویژگی سازماندهی می‌شود. عملیات ادغام[49] پس از هموارسازی نتایج انجام می شود، نقش آن ادغام ویژگی‌های مشابه معنایی در یکی است. به عبارت دیگر، ادغام تعداد پارامترها را کاهش می‌دهد و شبکه را از نظر محاسباتی کم هزینه می‌کند. Max Pooling یک عملیات ادغام معمولی است که با استخراج وصله‌ها (پچ‌ها)[50] از نقشه‌های ویژگی، تعیین حداکثر مقدار در هر پچ و سپس حذف تمام مقادیر دیگر ادامه می‌یابد. در نهایت، پس از تبدیل ماتریس‌های ورودی به یک بردار تک

[45] multilayer perceptron
[46] recurrent neural networks
[47] convolutional neural networks
[48] space-temporal structures
[49] pooling
[50] patches

بعدی، ویژگی‌ها توسط شبکه‌ای از لایه‌های کاملاً متصل مشابه شبکه عمیق پیش‌خور ذکر شده نگاشت می‌شوند تا خروجی نهایی، احتمالات یک ویژگی معین متعلق به یک کلاس معین به دست آید. خروجی شبکه عصبی کاملاً متصل به تابع فعال سازی متفاوت دیگری برای انجام طبقه بندی یا رگرسیون بر اساس متغیر خروجی تغذیه می‌شود. CNNها با موفقیت در تشخیص بصری و گفتار، پردازش زبان طبیعی و کارهای طبقه بندی مختلف استفاده شده‌اند (Yamashita et al. 2018).

مثال کاربردی: در پژوهشی .Bobbo et al (2021) روش‌های مختلف یادگیری ماشین برای پیش‌بینی وضعیت سلامت پستان بر اساس تعداد سلول‌های بدنی در گاوهای شیری را مقایسه کردند. آن‌ها هشت روش مختلف یادگیری ماشین یعنی تحلیل متمایز خطی[۵۱]، مدل خطی تعمیم‌یافته با تابع پیوند لاجیت[۵۲]، بیزهای ساده[۵۳]، درختان طبقه‌بندی و رگرسیون[۵۴]، K نزدیکترین همسایگان، ماشین‌های بردار پشتیبان، جنگل تصادفی و شبکه عصبی را برای پیش‌بینی وضعیت سلامت پستان گاوها، بر اساس تعداد سلول‌های سوماتیک مقایسه کردند. آن‌ها یک مجموعه داده ۱۸۴۴۲ رکوردی را به طور تصادفی به دو زیرمجموعه تقسیم کردند: ۸۰ درصد داده‌ها (۱۴۷۵۵ رکورد) برای آموزش و آزمایش مدل‌ها استفاده شد، در حالی که ۲۰ درصد باقی مانده از داده‌ها از ساخت مدل حذف شدند و به عنوان اعتبار سنجی خارجی نگهداری شدند. مجموعه انتخاب ویژگی بازگشتی با استفاده از یک اعتبارسنجی متقابل ده برابری که ۱۰۰ بار با روش RF تکرار می‌شود، برای انتخاب خودکار زیرمجموعه‌ای از پیش‌بینی‌کننده‌ترین ویژگی‌ها برای شناسایی مقرون‌به‌صرفه‌ترین مدل با بهترین عملکرد استفاده کردند. برای آموزش و آزمایش رابطه بین پیامد (0/1 بر اساس سطوح SCC در TD n + 1) و ویژگی‌ها (داده‌های گاو و شیر TD n) یک اعتبارسنجی متقابل ده برابری که ۱۰۰ بار تکرار شد، استفاده نمودند. طبقه بندی اجازه می‌دهد تا کلاس‌های خروجی نامتعادل را در نظر بگیرد و درصد نمونه‌ها را برای هر کلاس هدف حفظ کند. مجموعه داده اصلی آموزش/تست (n = 14755) را به طور تصادفی به ۱۰ زیر مجموعه با اندازه مساوی تقسیم کردند. مدل‌های پیش‌بینی را بر روی ۹ مورد از این زیرمجموعه‌ها آموزش دادند و آخرین زیرمجموعه را به عنوان مجموعه آزمون برای ارزیابی عملکرد روش‌ها در پیش‌بینی نتیجه استفاده کردند. هر اعتبارسنجی متقابل ده برابری را ۱۰۰ بار تکرار کردند. در مجموع ۱۰۰۰ تکرار انجام دادند و سپس میانگین ۱۰۰ مقدار دقت اعتبارسنجی متقابل ده برابری برای بدست آوردن دقت نهایی هر روش را به دست آوردند. استانداردسازی داده‌ها (مرکز و مقیاس) را در اعتبارسنجی متقابل انجام دادند. تجزیه و تحلیل داده‌ها را با استفاده از بسته‌های Caret v.6.0-8639و Tidyverse v. 1.3.140 نرم افزار R v.4.0.5 انجام دادند. دقت پیش‌بینی همه روش‌ها بالای ۷۵ درصد بود. با توجه به معیارهای مختلف، روش‌های شبکه عصبی، جنگل تصادفی و خطی بهترین عملکرد را در پیش‌بینی کلاس‌های سلامت پستان در یک روز آزمایشی معین (سالم یا ماستیتیک با توجه به تعداد سلول‌های بدنی زیر یا بالاتر از آستانه از پیش تعریف‌شده ۲۰۰۰۰۰ سلول در میلی‌لیتر) بر روی صفات شیر گاو ثبت شده در روز آزمون قبلی داشتند. یافته‌های آن‌ها الگوریتم‌های یادگیری

[51] linear discriminant analysis
[52] generalized linear model with logit link function
[53] naïve bayes
[54] classification and regression trees

ماشین را به عنوان ابزاری امیدوارکننده برای بهبود تصمیم‌گیری برای کشاورزان پیشنهاد کرد. تجزیه و تحلیل یادگیری ماشین روش‌های نظارت را بهبود می‌بخشد و به کشاورزان کمک می‌کند تا گاوهایی را که احتمالاً در روز آزمایش بعدی تعداد سلول‌های سوماتیک بالایی دارند، از قبل شناسایی کنند.

نتیجه‌گیری: این مطالعه نشان می‌دهد که استفاده از روش‌های یادگیری ماشین در حوزه ژنتیک و اصلاح نژاد و بیوتکنولوژی دام و طیور رو به افزایش است و با این افزایش کیفیت روش‌های یادگیری ماشین استفاده شده نیز رو به بهبود است. لذا، یادگیری ماشین می‌تواند در توسعه پایدار دامپروری و ارائه مزایایی مانند افزایش بهره‌وری در این حوزه نقش مهم و پررنگی را ایفا کند. بنابراین، این پژوهش توصیه می‌کند که استفاده از روش‌ها و الگوریتم‌های یادگیری ماشین در بین فعالان حوزه ژنتیک و اصلاح نژاد و بیوتکنولوژی دام و طیور ترویج شود تا زودتر و دقیق‌تر مشکلات را شناسایی و پیش بینی کنند و از بروز مشکلات و ضررهای اقتصادی جلوگیری نمایند.

References

Belyadi H, Haghighat A (2021). Machine learning guide for oil and gas using Python: A step-by-step breakdown with data, algorithms, codes, and applications. Gulf Prof Pub, pp.169–295.

Ben-Hur A, Horn D, Siegelmann HT, Vapnik V (2001) Support vector clustering. J Mach Learn Res 2, 125-137.

Bobbo T, Biffani S, Taccioli C, et al. (2021) Comparison of machine learning methods to predict udder health status based on somatic cell counts in dairy cows. Sci Rep 11, e13642.

Breiman L (1996) Bagging predictors. Mach Learn 24, 123-140.

Breiman L (2001) Random forests. Mach Learn 45(1), 5-32.

Brownlee J (2016) Bagging and random forest ensemble algorithms for machine learning. Mach Learn Algorithms 133.

Bühlmann P (2012) Bagging, boosting and ensemble methods. Handb Comput Statistics Concepts methods, 985-1022.

Cabrera VE, Barrientos-Blanco JA, Delgado H, Fadul-Pacheco L (2020) Symposium review: Real-time continuous decision making using big data on dairy farms. J Dairy Sci 103, e3856e3866.

Chafai N, Hayah I, Houaga I, Badaoui B (2023) A review of machine learning models applied to genomic prediction in animal breeding. Front Genet 14, e1150596.

Choudhary R, Gianey HK (2017) Comprehensive review on supervised machine learning algorithms. 2017 International Conference on Machine Learning and Data Science (MLDS). IEEE, 37-43.

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20, 273-297.

Crisci C, Ghattas B, Perera G (2012) A review of supervised machine learning algorithms and their applications to ecological data. Ecol Model 240, 113-122.

Crossa J, Montesinos-Lopez OA, Costa-Neto G, et al. (2024) Machine learning algorithms translate big data into predictive breeding accuracy. Trends Plant Sci 24, S1360-1385.

Defazio A, Bach F, Lacoste-Julien S (2014) SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. Adv Neural Inf Process Syst 27.

Ericsson Unnerstad H, Lindberg A, Persson Waller K, et al. (2009). Microbial aetiology of acute clinical mastitis and agent-specific risk factors. Vet Microbiol 137, e90e97.

Exterkate P, Groenen PJF, Heij C, van Dijk D (2016) Nonlinear forecasting with many predictors using kernel ridge regression. Int J Forecast 32(3), 736-53.

Fadul-Pacheco L, Delgado H, Cabrera VE (2021) Exploring machine learning algorithms for early prediction of clinical mastitis. Int Dairy J 119, e105051.

Forni S, Aguilar I, Misztal I (2011) Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genet Sel Evol 43, e1.

Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. ICML: Proceedings of the Thirteenth International Conference on Machine Learning 96, 148-156.

Ghotbaldini H, Mohammadabadi MR, Nezamabadi-pour H, et al. (2019) Predicting breeding value of body weight at 6-month age using Artificial Neural Networks in Kermani sheep breed. Acta Scientiarum Anim Sci 41, e45282.

Gianola D, de los Campos G (2018) Inferring genetic values for quantitative traits with regression models. Genetics 208(3), 1391-1404.

Gianola D, Okut H, Weigel KA, Rosa GJ (2011) Predicting complex quantitative traits with bayesian neural networks: A case study with Jersey cows and wheat. BMC Genet 12, 87-14.

González-Camacho JM, Ornella L, Pérez-Rodríguez P, et al. (2018) Applications of machine learning methods to genomic selection in breeding wheat for rust resistance. Plant Genome 11(2), e170104.

Gunn SR (1998) Support vector machines for classification and regression. ISIS Tech Rep 14(1), 5-16.

Jiang T, Gradus JL, Rosellini AJ (2020) Supervised machine learning: A brief primer. Behav Ther 51(5), 675-687.

Kingsford C, Salzberg SL (2008) What are decision trees? Nat Biotechnol 26(9), 1011-1013.

Kotlarz K, Mielczarek M, Biecek P et al. (2024) An Explainable Deep Learning Classifier of Bovine Mastitis Based on Whole-Genome Sequence Data—Circumventing the p >> n Problem. Int J Mol Sci 25, e4715.

LaValley MP (2008) Logistic regression. Circulation 117(18), 2395-2399.

Li X, Chen X, Wang Q, et al. (2024) Integrating bioinformatics and machine learning for genomic prediction in chickens. Genes 15, e690.

Madsen P, Jensen J, Labouriau R, et al. (2014) DMU-A Package for analyzing multivariate mixed models in quantitative genetics and genomics. In: Proceedings of the 10th World Congress of genetics applied to livestock production. August 17-22, Canada.

Mahesh B (2020) Machine learning algorithms-a review. Int J Sci Res 9(1), 381-386.

Manton JH, Amblard PO (2014) A primer on reproducing kernel Hilbert spaces. Available at: http://arxiv.org/abs/1408.0952 (Accessed June 18, 2019).

Maulud D, Abdulazeez AM (2020) A review on linear regression comprehensive in machine learning. J Appl Sci Technol Trends 1(4), 140-147.

Mohammadabadi M, Kheyrodin H, Afanasenko V, et al. (2024) The role of artificial intelligence in genomics. Agric Biotechnol J 16 (2), 195-279.

Montesinos-López OA, Montesinos-López A, Pérez-Rodríguez P, et al. (2021) A review of deep learning applications for genomic selection. BMC Genomics 22, 19-23.

Müller AC, Guido S (2017) Introduction to machine learning with Python: a guide for data scientists. O'Reilly Media, Inc: Sebastopol.

Nasteski V (2017) An overview of the supervised machine learning methods. Horizons B 4, 51-62.

Nayeri S, Sargolzaei M, Tulpan D (2019) A review of traditional and machine learning methods applied to animal breeding. Animal Health Res Rev 20(1), 31-46.

Nick TG, Campbell KM (2007) Logistic regression. Top Biostat 404, 273–301.

Pedregosa F, Varoquaux G, Gramfort A, et al. (2011) Scikit-Learn: Machine Learning in Python. J Mach Learn Res 12, 2825-2830.

Pour Hamidi S, Mohammadabadi MR, Asadi Foozi M, Nezamabadi-pour H (2017) Prediction of breeding values for the milk production trait in Iranian Holstein cows applying artificial neural networks. J Livestock Sci Technol 5(2), 53-61.

Shi S, Li X, Fang L, et al. (2021) Genomic prediction using Bayesian regression models with global-local prior. Front Genet 12, e628205.

Shrestha DL, Solomatine DP (2006) Experiments with AdaBoost.RT, an improved boosting scheme for regression. Neural Comput 18(7), 1678-1710.

Soloshenkov AD, Soloshenkova EA, Semina MT (2024) Artificial intelligence and classical methods in animal genetics and breeding. Russ J Genet 60(7), 843-856.

Vieira S, Pinaya WHL, Garcia-Dias R, Mechelli A (2020) Deep neural networks, in Machine learning (Academic Press), 157-172.

Wang X, Shi S, Wang G, et al. (2022) Using machine learning to improve the accuracy of genomic prediction of reproduction traits in pigs. J Anim Sci Biotechnol 13, e60.

Zhang G, Dai Z, Dai X (2020) C-RNNCrispr: prediction of CRISPR/Cas9 sgRNA activity using convolutional and recurrent neural networks. Comput Struct Biotechnol J 18, 344-354.

Zhang Z (2016) Introduction to machine learning: K-Nearest neighbors. Ann Transl Med 4(11), e218.