پیش‌بینی کیفیت شیر با روش یادگیری ماشین CatBoost و الگوریتم ژنتیک

امیرحسین اسمعیل پور

دانشجوی دکتری، گروه مهندسی صنایع، دانشکده فنی، دانشگاه خوارزمی، تهران، ایران

مریم عاملی

استادیار مهندسی صنایع، گروه مهندسی صنایع، دانشکده فنی، دانشگاه خوارزمی، تهران، ایران

چکیده

شیر یکی از کالاهای مهم در زنجیره غذایی در سبد خانوار است. از این رو کیفیت شیر از اهمیت ویژه‌ای برخوردار است زیرا تأثیر مستقیمی بر سلامت مصرف‌کنندگان و عملکرد صنایع لبنی دارد. در این پژوهش، از الگوریتم CatBoost به همراه تنظیم بهینه هایپرپارامترها با استفاده از الگوریتم ژنتیک، برای طبقه‌بندی کیفیت شیر به سه دسته "بالا"، "متوسط" و "ضعیف" استفاده شده است. همچنین، جهت افزایش تفسیرپذیری مدل از ابزار SHAP استفاده شده است. نتایج تحلیل SHAP نشان می‌دهد که ویژگی چربی و بوی شیر خام بیشترین تأثیر را بر خروجی مدل دارد. عملکرد مدل با معیارهای ارزیابی به‌دست‌آمده به ترتیب عبارت است از: دقت (Precision) برابر با ۰.۹۸۶۷، یادآوری (Recall) برابر با ۰.۹۹۱۵، امتیاز-F۱ (F۱-score) برابر با ۰.۹۸۸۹ و صحت (Accuracy) برابر با ۰.۹۹۰۶. این نتایج حاکی از دقت و قابلیت اطمینان بالای مدل است که می‌تواند به عنوان یک راهکار سریع، مقرون‌به‌صرفه و مطمئن در بهبود فرآیندهای کنترل کیفیت صنایع لبنی مورد استفاده قرار گیرد.

واژگان کلیدی: پیشبینی کیفیت شیر، یادگیری ماشین، الگوریتم ژنتیک، کنترل کیفیت

مقدمه

در دنیای امروز، شیر به عنوان یکی از منابع اصلی تغذیه و یکی از کالاهای حیاتی در زنجیره غذایی، نقشی بی‌بدیل در سلامت جامعه و توسعه اقتصادی دارد. کیفیت این ماده غذایی تأثیر مستقیمی بر سلامت مصرف‌کنندگان و عملکرد صنایع فرآوری لبنی دارد؛ بنابراین، پیش‌بینی سریع و دقیق تغییرات کیفیت شیر امری ضروری است. هرچه بتوان در اسرع وقت از ناهنجاری‌ها و کاهش کیفیت آگاه شد، می‌توان با اجرای اقدامات اصلاحی به موقع، از بروز خسارات جدی اقتصادی و سلامتی جلوگیری نمود (دردشتی, کریم, بکایی, & لاری, ۲۰۰۰).

با پیشرفت فناوری‌های نوین در حوزه داده‌کاوی، هوش مصنوعی و یادگیری ماشین به ابزارهای قدرتمندی برای تحلیل داده‌های پیچیده تبدیل شده‌اند. این روش‌ها با استخراج الگوهای پنهان و شناسایی روندهای غیرخطی، امکان پیش‌بینی به‌موقع تغییرات کیفیت را فراهم می‌کنند و از ورود محصولات نامرغوب به بازار جلوگیری می‌شود. به‌کارگیری یادگیری ماشین در تحلیل کیفیت شیر، امکان شناسایی جامع روندهای پنهان و عوامل مؤثر را فراهم می‌کند. با توجه به اهمیت بهبود فرآیندهای کنترل کیفیت در صنعت لبنیات و نیاز به تصمیم‌گیری‌های سریع و دقیق، استفاده از فناوری‌های هوشمند می‌تواند زمینه‌ساز تحولاتی در این حوزه شود. این رویکرد نوین، با به کارگیری الگوریتم‌های پیشرفته یادگیری ماشین، امکان پردازش سریع داده‌ها و ارائه پیش‌بینی‌های دقیق را فراهم می‌آورد که در نهایت به افزایش بهره‌وری، کاهش ضایعات و ارتقای سطح کیفیت محصولات لبنی منجر خواهد شد (Xiao, Xia, & Tian, 2019) (سعادتفر, ۱۳۹۵).

پیش‌بینی کیفیت شیر با استفاده از روش‌های یادگیری ماشین در سال‌های اخیر مورد توجه گسترده‌ای قرار گرفته است. این روش‌ها با تحلیل ویژگی‌های مختلف شیر، امکان تشخیص سریع و دقیق کیفیت آن را فراهم می‌کنند.

در پژوهشی، سیستم MilkSafe طراحی شده که با کمک سخت‌افزار و مدل‌های یادگیری ماشین، کیفیت شیر را ارزیابی می‌کند. داده‌های مربوط به ویژگی‌های شیر شامل pH، دما، کدورت و رنگ توسط حسگرها جمع‌آوری شده و به مدل یادگیری ماشین وارد می‌شود. بر اساس این ویژگی‌ها، کیفیت شیر در سه سطح پایین، متوسط و بالا طبقه‌بندی می‌شود. در این سامانه از میکروکنترلر Arduino UNO برای جمع‌آوری اطلاعات از حسگرها استفاده شده و خروجی‌ها در محیط Arduino IDE نمایش داده می‌شوند. مدل با استفاده از این داده‌ها آموزش داده شده و الگوریتم‌های مختلف از جمله Naive Bayes، Random Forest، KNN و Logistic Regression مورد آزمایش قرار گرفته‌اند که در این میان، Random Forest بالاترین دقت را ارائه داده است. مدل پیشنهادی با استفاده از چهار ویژگی ورودی به دقت ۹۸,۲۷٪ دست یافته و یک سیستم خودکار و قابل اعتماد برای ارزیابی کیفیت شیر ارائه کرده است (H. V, S, Jha, & S, 2023).

این مطالعه به بررسی عوامل مؤثر بر کیفیت شیر پرداخته و ۹ متغیر شامل pH، دما، طعم، بو، چربی، کدورت، رنگ و درجه‌بندی را مورد تجزیه‌وتحلیل قرار داده است. با استفاده از روش تحلیل مؤلفه‌های اصلی (PCA) مشخص شد که دما و رنگ بیشترین تأثیر را بر کیفیت شیر دارند و بیش از ۹۵٪ از واریانس در مؤلفه‌های اصلی اول و دوم (PCA-۱ و PCA-۲) را توضیح می‌دهند. همچنین، نمونه‌های شیر به سه سطح پایین، متوسط و بالا دسته‌بندی شده و برای طبقه‌بندی آن‌ها از الگوریتم شبکه عصبی مصنوعی

(ANN) استفاده شده است. نتایج نشان داد که مدل ANN توانست نمونه‌های شیر را با دقت ۰,۹۹۸۸ طبقه‌بندی کند و در مقایسه با روش‌های دیگر از دقت و پایداری بالاتری برخوردار بود. علاوه بر این، استفاده از الگوریتم‌های پیشرفته‌تر یادگیری عمیق می‌تواند به بهبود نتایج کمک کند (Kumari, Gourisaria, Das, & Banik, 2023).

شیر به عنوان یکی از مواد غذایی اصلی انسان، در معرض تقلب قرار دارد که می‌تواند سلامت مصرف‌کنندگان را به خطر بیندازد و به اعتبار صنعت لبنیات آسیب بزند. یکی از رایج‌ترین روش‌های تقلب، افزودن مواد شیمیایی، به‌ویژه آب، به شیر است. روش‌های سنتی ارزیابی کیفیت شیر مانند ارزیابی حسی، آنالیز شیمیایی و بررسی‌های میکروبیولوژیکی در تشخیص تقلب‌های با غلظت کم چندان مؤثر نیستند. ترکیب شیمیایی شیر پیچیده است و شناسایی مواد تقلبی در این ماتریس پیچیده نیازمند تکنیک‌های تحلیلی پیشرفته است. این پژوهش، رویکردی نوین برای مقابله با تقلب در شیر را بررسی می‌کند که در آن، یادگیری ماشین با فناوری حسگرها ترکیب شده است. سیستم پیشنهادی شامل سخت‌افزارهایی مانند Arduino Uno و حسگرهای مختلف (pH، کدورت، دما) برای جمع‌آوری داده‌های مرتبط با کیفیت شیر است. داده‌ها پس از آزمایش هشت نمونه مختلف شیر شامل شیر اسکیم، هموژنیزه، استاندارد با غلظت‌های چربی مختلف گردآوری شده و سپس با استفاده از الگوریتم‌های یادگیری ماشین از جمله ماشین بردار پشتیبان (SVM)، طبقه‌بندی‌کننده Adaboost و جنگل تصادفی (Random Forest) پردازش و تحلیل شده‌اند تا کیفیت شیر پیش‌بینی شود (Sunithamani, Muralidhar, Anne, & Sruthi, 2024).

به‌طور کلی، ادغام روش‌های یادگیری ماشین در فرآیندهای تولید و کنترل کیفیت شیر می‌تواند به بهبود کیفیت محصولات لبنی و افزایش رضایت مصرف‌کنندگان کمک کند. با توجه به اهمیت بهبود سیستم‌های کنترل کیفیت در صنعت لبنی و نیاز به کاهش هزینه‌های تولید، استفاده از رویکردهای نوین مبتنی بر هوش مصنوعی و ابزارهای تحلیل ویژگی امری ضروری به نظر می‌رسد. این مقاله سعی دارد با بهره‌گیری از یک مدل CatBoost به همراه الگوریتم ژنتیک برای تنظیم هایپرپارامترها، به پیش‌بینی کیفیت شیر پرداخته و به بررسی دقیق تأثیر ویژگی‌های مختلف مانند pH، دما، طعم، بو، میزان چربی، کدورت و رنگ بر کیفیت شیر بپردازد. نتایج حاصل از این پژوهش می‌تواند به جهت بهبود فرآیندهای کنترل کیفیت در صنایع لبنی و ارائه راهکارهای نوین در این حوزه مورد استفاده قرار گیرد.

روش تحقیق

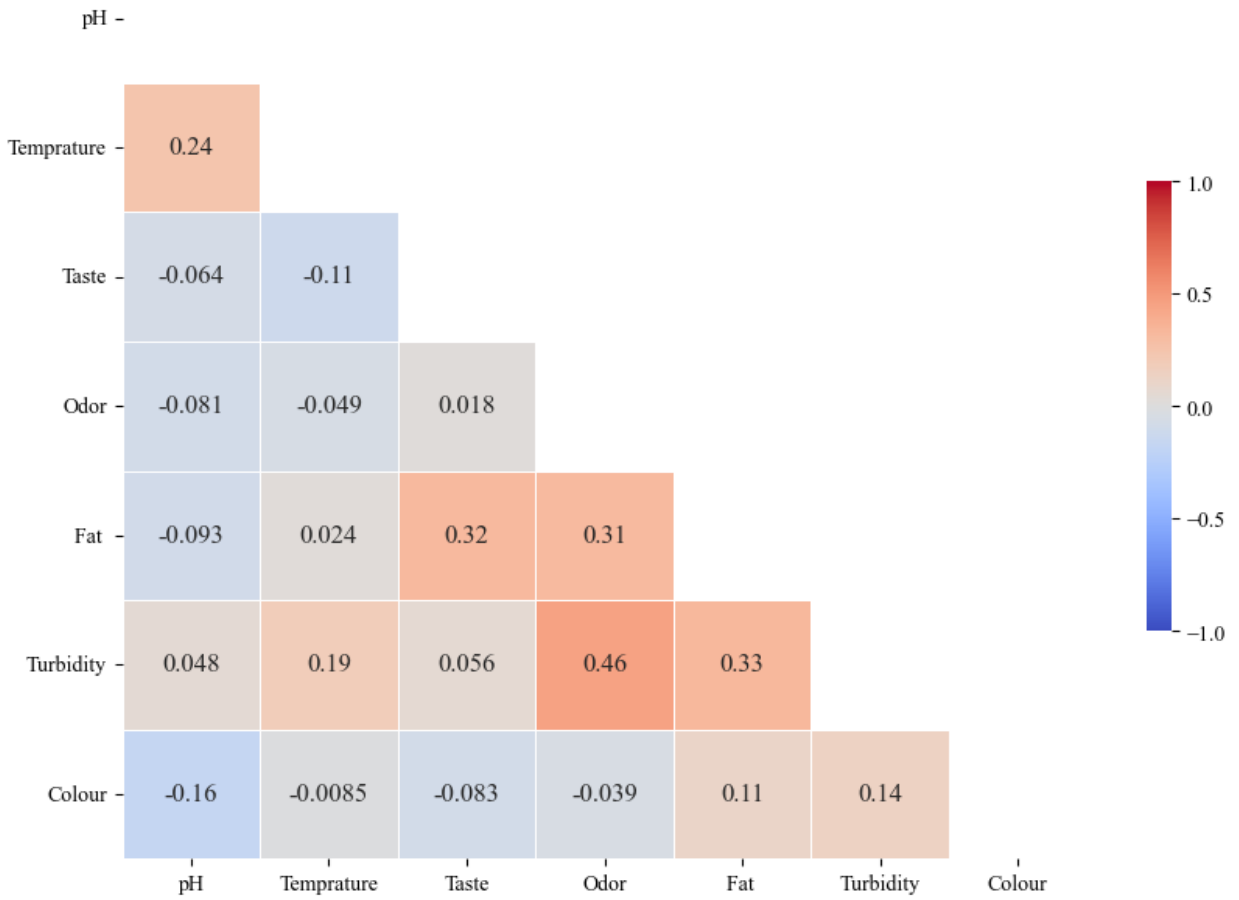
مجموعه داده سیستم پیشنهادی از مخزن Kaggle جمع‌آوری شده است (GNV, 2021). این مجموعه داده شامل ۷ ویژگی مستقل است (همانطور که در جدول ۱ نشان داده شده) که از آن‌ها برای تحلیل و پیش‌بینی کیفیت شیر استفاده می‌شود. هدف (برچسب) شیر، داده‌ای طبقه‌بندی شده است که در پروژه "پیش‌بینی کیفیت شیر با استفاده از یادگیری ماشین" به سه کلاس مختلف شامل Low (ضعیف)، Medium (متوسط) و High (بالا) تقسیم شده است. تعداد کل رکوردهای موجود در این مجموعه داده ۱۰۶۰ ردیف و ۸ ستون می‌باشد که از میان این ویژگی‌ها، ۷ ویژگی طبقه‌ای و ۱ ویژگی عددی هستند. در این پژوهش از پایتون که یک زبان برنامه‌نویسی قدرتمند و پرکاربرد در یادگیری ماشین و علم داده است.

جدول ۱ آمار توصیفی مجموعه داده را برای ویژگی‌های مختلف مرتبط با کیفیت شیر نشان می‌دهد. Count تعداد کل نمونه‌ها را نمایش می‌دهد که برای همه ویژگی‌ها برابر با ۱۰۵۹ است. Mean مقدار میانگین هر ویژگی را نشان می‌دهد، مانند میانگین pH که برابر ۶.۶۳ است. Std انحراف معیار را مشخص می‌کند که میزان پراکندگی داده‌ها را نشان می‌دهد. Min و Max به ترتیب حداقل و حداکثر مقدار هر ویژگی را نمایش می‌دهند. چارکهای ۲۵٪، ۵۰٪ (میانه) و ۷۵٪ نیز برای درک توزیع داده‌ها ارائه شده‌اند. داده ها شامل pH، دما (Temperature)، طعم (Taste)، بو (Odor)، چربی (Fat)، کدورت (Turbidity)، رنگ (Colour) است.

جدول ۱. توصیفی از ویژگی‌های مختلف مجموعه داده شیر شامل میانگین(mean)، انحراف معیار(std)، حداقل(min)، حداکثر(max) و چارک‌های ۲۵٪، ۵۰٪ و ۷۵٪

	pH	*Temperature*	*Taste*	*Odor*	*Fat*	*Turbidity*	*Colour*
count	۱۰۵۹	۱۰۵۹	۱۰۵۹	۱۰۵۹	۱۰۵۹	۱۰۵۹	۱۰۵۹
mean	۶.۶۳۰۱۲۳	۴۴.۲۲۶۶۳	۰.۵۴۶۷۴۲	۰.۴۳۲۴۸۳	۰.۶۷۱۳۸۸	۰.۴۹۱۰۲۹	۲۵۱.۸۴۰۴
std	۱.۳۹۹۶۷۹	۱۰.۰۹۸۳۶	۰.۴۹۸۰۴۶	۰.۴۹۵۶۵۵	۰.۴۶۹۹۳	۰.۵۰۰۱۵۶	۴.۳۰۷۴۲۴
min	۳	۳۴	۰	۰	۰	۰	۲۴۰
۲۵٪	۶.۵	۳۸	۰	۰	۰	۰	۲۵۰
۵۰٪	۶.۷	۴۱	۱	۰	۱	۰	۲۵۵
۷۵٪	۶.۸	۴۵	۱	۱	۱	۱	۲۵۵
max	۹.۵	۹۰	۱	۱	۱	۱	۲۵۵

شکل ۱ تصویر یک ماتریس همبستگی را نمایش می‌دهد که میزان وابستگی میان ویژگی‌های مختلف کیفیت شیر را نشان می‌دهد. مقدارهای مثبت‌تر (رنگ‌های قرمز) بیانگر همبستگی مستقیم بین دو ویژگی هستند، درحالی‌که مقدارهای منفی‌تر (رنگ‌های آبی) نشان‌دهنده همبستگی معکوس‌اند. برای مثال، چربی (Fat) و کدورت (Turbidity) دارای همبستگی نسبتاً مثبت ۰.۳۳ هستند، درحالی‌که pH و رنگ (Colour) همبستگی منفی ۰.۱۶ دارند. این تحلیل می‌تواند درک بهتری از تأثیر ویژگی‌ها بر یکدیگر فراهم کند.

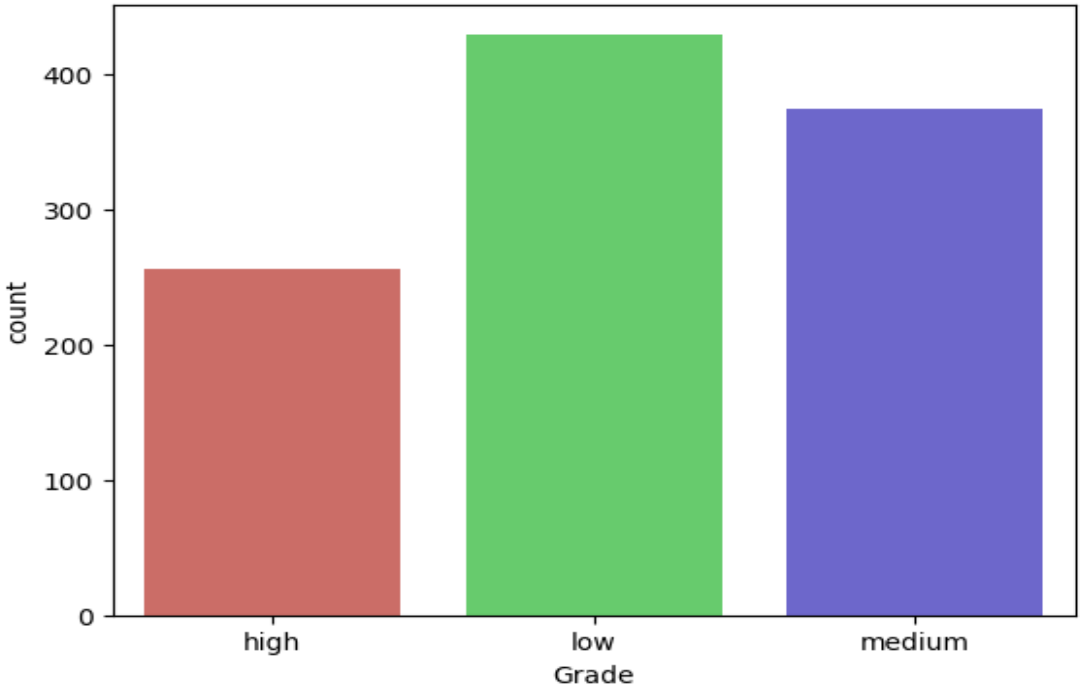


شکل ۱. ماتریس همبستگی ویژگی‌های کیفیت شیر، نمایش میزان وابستگی بین متغیرها

داده ها به دو دسته train و test با نسبت ۸۰٪ و ۲۰٪ به ترتیب تبدیل شده اند.

شکل ۲ نمودار توزیع برچسب‌های کیفیت شیر در مجموعه داده اولیه را نشان می‌دهد که نشان‌دهنده عدم توازن بین کلاس‌ها است. به دلیل عدم توازن در توزیع کلاس‌های داده، از تکنیک SMOTE برای افزایش نمونه‌های کلاس‌های کمتر و بهبود عملکرد مدل استفاده شد. SMOTE (Synthetic Minority Over-sampling Technique) یک روش برای افزایش نمونه‌های

کلاس‌های کمتر در داده‌های نامتوازن با تولید داده‌های مصنوعی است (Joloudari, Marefat, Nematollahi, Oyelere, & Hussain, 2023).



شکل ۲. نمودار توزیع برچسب‌های کیفیت شیر در مجموعه داده اولیه

CatBoost یک الگوریتم گرادیان بوستینگ است که به‌ویژه برای مسائل طبقه‌بندی و رگرسیون مناسب بوده و از ویژگی‌های دسته‌ای (categorical features) به شکلی هوشمندانه بهره می‌برد. این الگوریتم با کاهش احتمال بیش برازش (overfitting) و افزایش دقت پیش‌بینی، به‌عنوان یکی از ابزارهای قدرتمند در یادگیری ماشین مطرح شده است. از ویژگی‌های بارز CatBoost می‌توان به قابلیت خودکارسازی پردازش ویژگی‌های دسته‌ای و سازگاری بالا با داده‌های پیچیده اشاره کرد (Wang & Qian, ۲۰۲۳).

تنظیم بهینه هایپرپارامترها یکی از مراحل کلیدی در بهبود عملکرد مدل‌های یادگیری ماشین است. الگوریتم ژنتیک یک روش بهینه‌سازی مبتنی بر اصول انتخاب طبیعی است که با تولید جمعیتی از راه‌حل‌های تصادفی (هادی‌ها یا کروموزوم‌ها)، از طریق فرآیندهایی مانند انتخاب، تقاطع و جهش، به مرور زمان به سمت بهترین ترکیب‌های ممکن حرکت می‌کند (Brzęk, Probierz, & Kozak, 2025). در کاربردهای مربوط به CatBoost، الگوریتم ژنتیک امکان تنظیم خودکار هایپرپارامترهایی مانند تعداد تکرارها، نرخ یادگیری، عمق درخت‌ها، ضریب منظم‌سازی و سایر پارامترهای مؤثر را فراهم می‌کند؛ بدین ترتیب مدل با تنظیم بهینه پارامترها، به دقت و کارایی بالاتری در پیش‌بینی دست می‌یابد. جدول ۲ هایپرپارامترهای CatBoost و بهترین هایپرپارامتر

انتخابی توسط الگوریتم ژنتیک را نشان می‌دهد. در این مطالعه، از الگوریتم ژنتیک برای تنظیم هایپرپارامترهای CatBoost استفاده شده است. تنظیمات اصلی این الگوریتم به شرح زیر است:

اندازه جمعیت ۲۰ :(Population Size)

تعداد نسل‌ها ۱۰ :(Generations)

نرخ جهش ۰.۲ :(Mutation Rate)

نرخ تقاطع ۰.۸ :(Crossover Rate)

روش انتخاب (Selection Method) : انتخاب تورنمنتی (Tournament Selection)

معیار ارزیابی (Fitness Function) : دقت مدل (Accuracy)

جدول ۲. هایپرپارامترهای CatBoost و بهترین هایپرپارامتر انتخابی توسط الگوریتم ژنتیک

هایپرپارامترها	توصیف کوتاه	رنج تست شده	بهترین مقدار
iterations	تعداد تکرارهای بوستینگ	۵۰۰ – ۱۵۰۰	۵۹۵
learning_rate	نرخ یادگیری برای فرآیند بوستینگ	۰.۰۱ – ۰.۳	۰.۲۳۸۷
depth	عمق درخت‌های مدل	۳ – ۱۰	۸
l۲_leaf_reg	ضریب منظم‌سازی L^2 برای وزن‌های برگ	۱ – ۱۰	۴.۹۰۶۹
bagging_temperature	کنترل شدت بیگینگ بیزی	۰ – ۱	۰.۳۷۲۳
random_strength	میزان تصادفی بودن در ارزیابی ویژگی‌ها	۰ – ۱	۰.۵۲۴۱

در ارزیابی عملکرد یک مدل طبقه‌بندی، چهار معیار مهم مورد استفاده قرار می‌گیرد که هر یک معنا و کاربرد خاص خود را دارند. در ادامه، توضیح هر یک از این معیارها به همراه فرمول محاسبه آورده شده است.

به‌منظور ارزیابی دقت و کارایی الگوریتم‌ها در پیش‌بینی، از جمله شاخص‌هایی که مورد بررسی قرار گرفته‌اند شامل صحت (Accuracy)، دقت(Precision)، بازخوانی (Recall) و امتیاز-F۱ (F۱-score$F1$)) هستند که محاسبه هر کدام به شرح زیر است:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \quad (۱)$$

$$Precision = \frac{tp}{tp + fp} \quad (۲)$$

$$Recall = \frac{tp}{tp + fn} \quad (۳$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (۴$$

tp = الگوریتم نمونه را در دسته مثبت طبقه‌بندی کرده و نمونه هم مثبت است

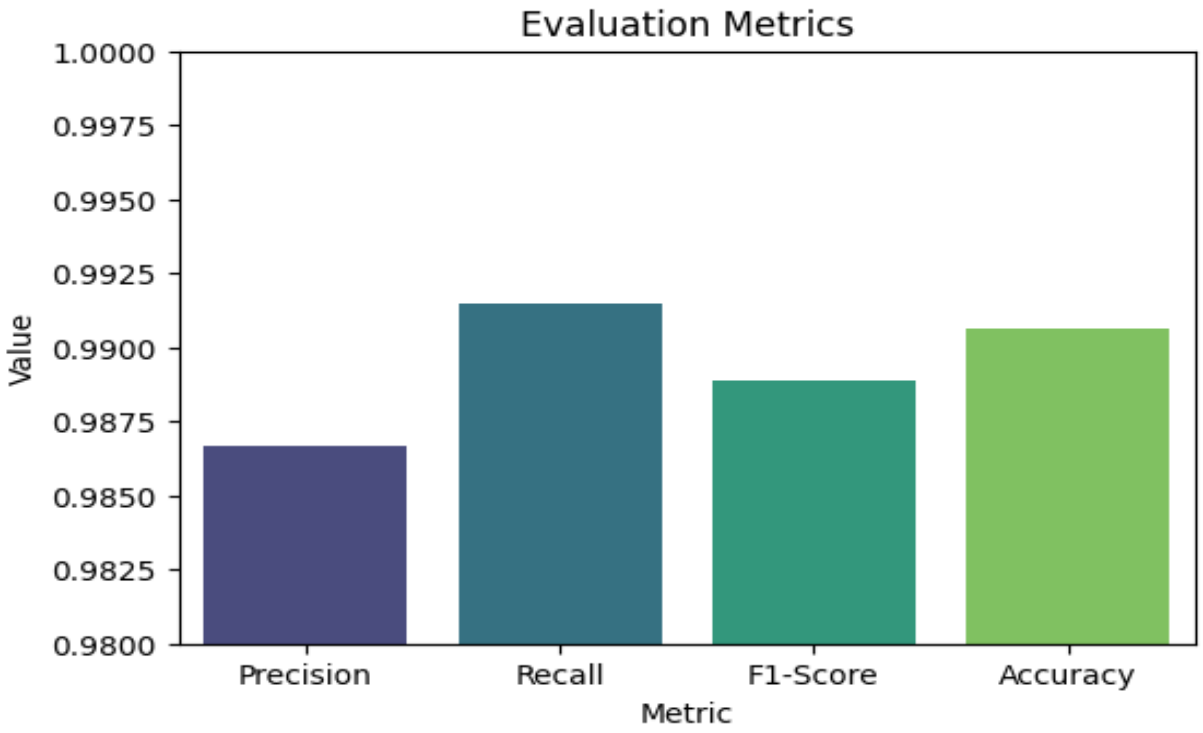
tn = الگوریتم نمونه را در دسته منفی طبقه‌بندی کرده و نمونه هم منفی است

fp = الگوریتم نمونه را در دسته مثبت طبقه‌بندی کرده اما نمونه منفی است

fn = الگوریتم نمونه را در دسته منفی طبقه‌بندی کرده اما نمونه مثبت است

یافته ها

شکل ۳ نمودار میانگین معیارهای ارزیابی مدل یادگیری ماشین در طبقه‌بندی کیفیت شیر را نشان می‌دهد. مقدار Precision نشان می‌دهد که چه نسبتی از نمونه‌های پیش‌بینی‌شده برای هر کلاس واقعاً صحیح هستند. مقدار Recall بیانگر میزان بازیابی صحیح نمونه‌های متعلق به هر کلاس از میان داده‌های واقعی است. F۱-Score ترکیبی از دقت و بازخوانی است که در داده‌های نامتوازن اهمیت بیشتری دارد. در نهایت، Accuracy عملکرد کلی مدل را بر اساس تمام کلاس‌ها ارزیابی می‌کند. نتایج نشان می‌دهند که مدل به Accuracy کلی ۹۹.۰۶٪ دست یافته است، و مقادیر بالای Precision، Recall و F۱-Score به ترتیب ۹۸.۶۷٪، ۹۹.۱۵٪ و ۹۸.۸۹٪ نشان‌دهنده عملکرد بسیار مطلوب مدل در پیش‌بینی کیفیت شیر هستند.



شکل ۳. نمودار مقادیر میانگین معیارهای ارزیابی مدل در پیش‌بینی کیفیت شیر

شکل ۴ تصویر ماتریس سردرگمی (Confusion Matrix) مدل یادگیری ماشین را برای پیش‌بینی کیفیت شیر نشان می‌دهد. در این ماتریس:

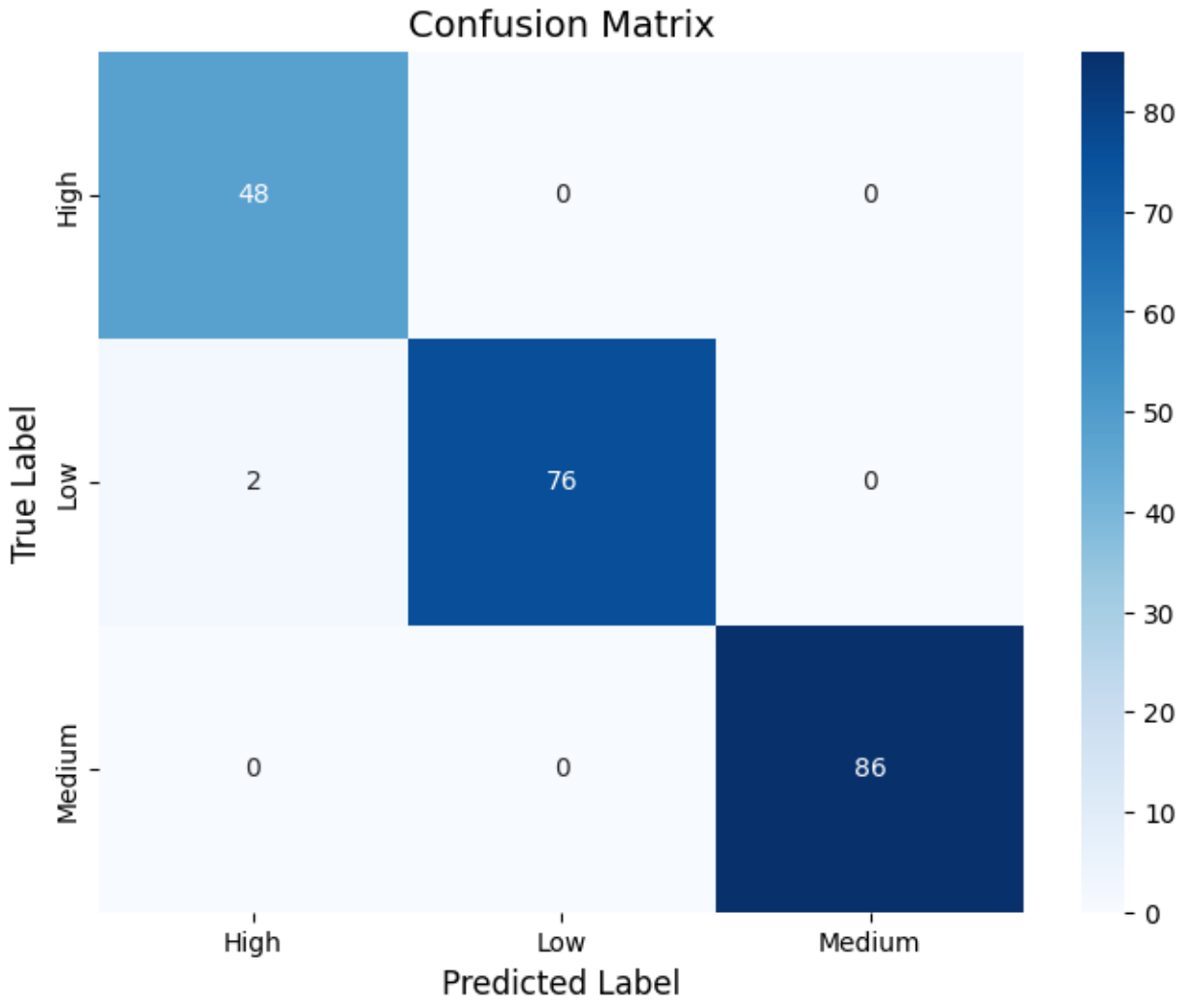
سطرها نشان‌دهنده کلاس‌های واقعی (True Label) هستند.

ستون‌ها نشان‌دهنده کلاس‌های پیش‌بینی‌شده (Predicted Label) توسط مدل هستند.

مقادیر روی قطر اصلی ماتریس تعداد نمونه‌هایی را نشان می‌دهند که مدل به‌درستی پیش‌بینی کرده است.

مقادیر خارج از قطر اصلی نشان‌دهنده نمونه‌هایی هستند که مدل به اشتباه در کلاس دیگری طبقه‌بندی کرده است.

مدل عملکرد بسیار دقیقی دارد، زیرا بیشتر پیش‌بینی‌ها روی قطر اصلی قرار دارند. به عنوان مثال، کلاس ضعیف ۷۶ (Low) نمونه صحیح پیش‌بینی شده و هیچ نمونه‌ای اشتباه طبقه‌بندی نشده است. تنها اشتباه مدل مربوط به ۲ نمونه از کلاس بالا (High) است که در کلاس ضعیف پیش‌بینی شده‌اند، که عملکرد بالای مدل را نشان می‌دهد.

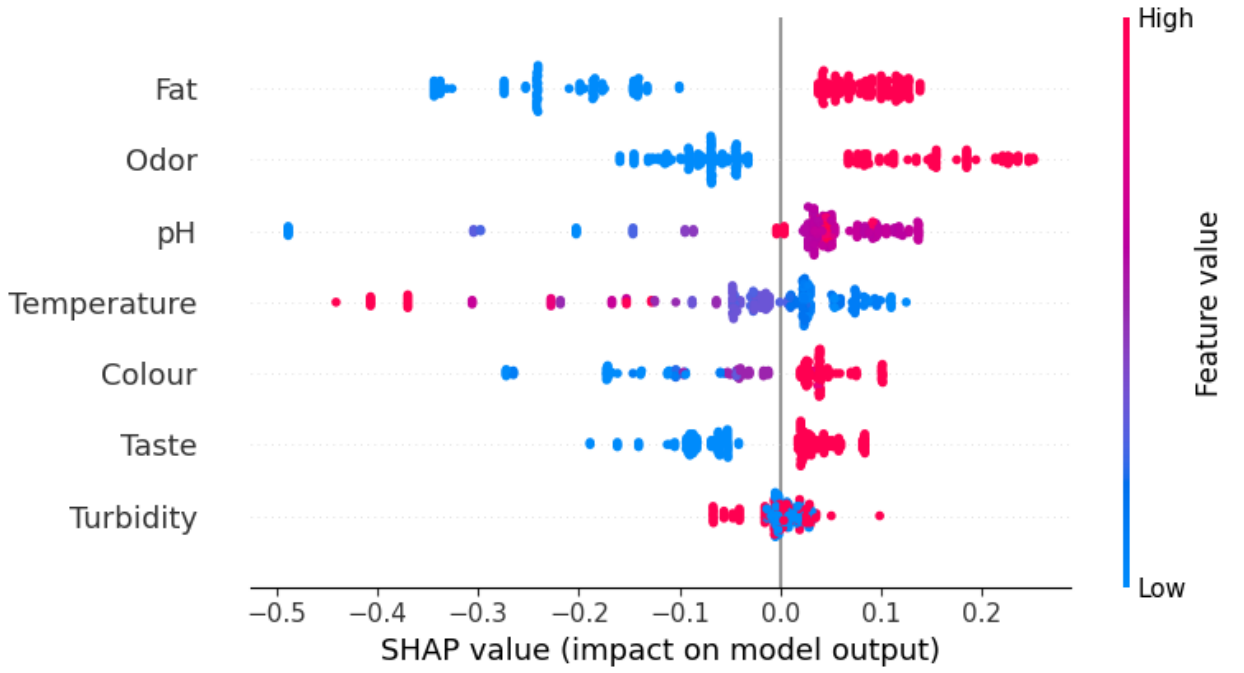


شکل ۴. ماتریس سردرگمی (Confusion Matrix) عملکرد مدل یادگیری ماشین را در پیش‌بینی کیفیت شیر

یکی از چالش‌های استفاده از مدل‌های پیچیده یادگیری ماشین، تفسیرپذیری آن‌هاست. SHAP ابزاری است که بر مبنای نظریه بازی‌ها توسعه یافته و به‌طور دقیق تأثیر هر ویژگی بر خروجی مدل را به صورت عددی و گرافیکی محاسبه می‌کند. استفاده از SHAP در کنار CatBoost و الگوریتم ژنتیک، به پژوهشگر کمک می‌کند تا به درک عمیق‌تری از دلایل پشت پیش‌بینی‌های مدل برسد و بتواند با ارائه توضیحات شفاف، از صحت و اعتبار نتایج اطمینان حاصل نماید (Hamilton & Papadopoulos, 2024).

شکل ۵ یک نمودار SHAP Summary Plotرا نشان می‌دهد که تأثیر ویژگی‌های مختلف بر خروجی مدل را بررسی می‌کند. محور افقی مقادیر SHAP را نمایش می‌دهد که نشان‌دهنده میزان تأثیر هر ویژگی بر پیش‌بینی مدل است. هر نقطه نشان‌دهنده یک نمونه از داده‌های ورودی است. رنگ نقاط نشان‌دهنده مقدار ویژگی مربوطه است، جایی که رنگ آبی مقادیر پایین و رنگ قرمز مقادیر بالا را نشان می‌دهد. این توزیع‌ها نشان می‌دهند که چگونه تغییر مقدار هر ویژگی باعث تغییر در پیش‌بینی مدل می‌شود. ویژگی‌ها بر اساس اهمیت آن‌ها در مدل مرتب شده‌اند، به طوری که "Fat" بیشترین تأثیر را در پیش‌بینی دارد. ویژگی‌هایی که

نقاط آن‌ها بیشتر در سمت راست قرار گرفته‌اند، تأثیر مثبتی بر کلاس "High" دارند، در حالی که نقاط در سمت چپ تأثیر منفی دارند. در این نمودار، متغیرهایی مانند "Fat" و "Odor" بیشترین تأثیر را در پیش‌بینی دارند، در حالی که "Temperature" و "Turbidity" تأثیر کمتری دارند.



شکل ۵. نمودار SHAP تأثیر ویژگی‌های مختلف بر مدل پیش‌بینی کیفیت شیر: مقادیر بالاتر ویژگی‌ها (قرمز) معمولاً تأثیر مثبتی بر طبقه "High" دارند، در حالی که مقادیر پایین (آبی) تأثیر بیشتری بر سایر طبقات دارند.

بحث و نتیجه‌گیری

با توجه به اهمیت کیفیت شیر در سلامت مصرف‌کنندگان و پایداری صنایع لبنی، این پژوهش یک چارچوب نوین برای پیش‌بینی کیفیت شیر با استفاده از یادگیری ماشین و هوش مصنوعی ارائه می‌دهد. در این مطالعه، از الگوریتم CatBoost به عنوان یک مدل قدرتمند یادگیری ماشین بهره گرفته شد که توانست با استفاده از ویژگی‌های مختلف فیزیکوشیمیایی شیر، به طبقه‌بندی دقیق نمونه‌ها در سه دسته "بالا"، "متوسط" و "ضعیف" دست یابد. تنظیم بهینه هایپرپارامترهای مدل با الگوریتم ژنتیک، عملکرد را بهبود بخشید و دقت پیش‌بینی را افزایش داد. علاوه بر این، استفاده از ابزار SHAP امکان تفسیر دقیق تأثیر هر ویژگی بر خروجی مدل را فراهم ساخت که به درک بهتر عوامل مؤثر در کیفیت شیر کمک می‌کند. نتایج به دست آمده، با صحت کلی بالای ۹۹.۰۶ درصد و مقادیر مطلوب دقت، یادآوری و امتیاز-F۱، نشان می‌دهد که رویکرد پیشنهادی می‌تواند به عنوان ابزاری کارآمد در نظارت و کنترل کیفیت محصولات لبنی به کار گرفته شود. در نهایت، این پژوهش امکان بهبود روش‌های پیش‌بینی کیفیت در صنایع غذایی را فراهم می‌کند، به‌ویژه با بهره‌گیری از داده‌های جامع‌تر و الگوریتم‌های پیشرفته یادگیری عمیق. توصیه می‌شود

تحقیقات آتی علاوه بر بهبود تنظیم خودکار هایپرپارامترها، به افزایش قابلیت تفسیرپذیری مدل‌ها و کاربرد آن‌ها در شرایط واقعی تولید نیز بپردازند تا بتوانند بهینه‌ترین راهکارهای نظارتی و کنترلی را ارائه دهند.

منابع

Brzęk, B., Probierz, B., & Kozak, J. (۲۰۲۵). Exploration-Driven Genetic Algorithms for Hyperparameter Optimisation in Deep Reinforcement Learning. *Applied Sciences, 15*(٤), ۲۰٦۷. Retrieved from https://www.mdpi.com/۲۰۷٦-۳٤۱۷/۱٥/٤/۲۰٦۷

Grading. Retrieved from: https://www.kaggle.com/datasets/prudhvignv/milk-grading GNV, P. (۲۰۲۱). *Milk*

H. V, T., S, S., Jha, S., & S, B. (۲۰۲۳, ٥-٦ May ۲۰۲۳). *MilkSafe: A Hardware-Enabled Milk Quality Prediction using Machine Learning.* Paper presented at the ۲۰۲۳ ۲nd International Conference on Vision Towards Emerging .Trends in Communication and Networking Technologies (ViTECoN)

Hamilton, R. I., & Papadopoulos, P. N. (۲۰۲٤). Using SHAP Values and Machine Learning to Understand Trends in the Transient Stability Limit. *IEEE Transactions on Power Systems, 39*(۱), ۱۳۸٤-۱۳۹۷. doi:۱۰,۱۱۰۹/TPWRS.۲۰۲۳,۳۲٤۸۹٤۱

Joloudari, J. H., Marefat, A., Nematollahi, M. A., Oyelere, S. S., & Hussain, S. (۲۰۲۳). Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks. *Applied Sciences, 13*(٦), ٤۰۰٦. Retrieved from https://www.mdpi.com/۲۰۷٦-۳٤۱۷/۱۳/٦/٤۰۰٦

Kumari, S., Gourisaria, M. K., Das, H., & Banik, D. (۲۰۲۳, ۲۸-۲۹ April ۲۰۲۳). *Deep Learning Based Approach for Milk Quality Prediction.* Paper presented at the ۲۰۲۳ ۱۱th International Conference on Emerging Trends in .Engineering & Technology - Signal and Information Processing (ICETET - SIP)

Sunithamani, S., Muralidhar, D., Anne, G., & Sruthi, C. N. (۲۰۲٤, ۱۲-۱٤ April ۲۰۲٤). *Milk quality prediction using Machine Learning integrated with Arduino.* Paper presented at the ۲۰۲٤ ۱۰th International Conference on .Communication and Signal Processing (ICCSP)

Wang, D., & Qian, H. (۲۰۲۳). CatBoost-Based Automatic Classification Study of River Network. *ISPRS International Journal of Geo-Information, 12*(۱۰), ٤۱٦. Retrieved from https://www.mdpi.com/۲۲۲۰-۹۹٦٤/۱۲/۱۰/٤۱٦

Xiao, L., Xia, K., & Tian, H. (۲۰۱۹, ۲۱-۲٤ Nov. ۲۰۱۹). *Research on Classification Model of Fermented Milk* International Conference on Intelligent *Quality Control Based on Data Mining.* Paper presented at the ۲۰۱۹ .Informatics and Biomedical Sciences (ICIIBMS)

دردشتی, د. ا. د., کریم, د. گ., بکایی, د. س., & لاری, د. م. ا. (۲۰۰۰). مطالعه کیفیت شیرهای تحویلی به کارخانه صنایع شیر ایران براساس اندازه گیری شاخصهای مختلف شیمیایی و شمارش کلی باکتریایی. مجله تحقیقات دامپزشکی *(Journal of Veterinary Research), 55*(3). Retrieved from https://jvr.ut.ac.ir/article_16746.html

سعادتفر, س. (۱۳۹۵). بررسی کیفیت میکروبی شیر خام استان گلستان به کمک شبکه های عصبی مصنوعی. Paper presented at the سومین کنگره علمی پژوهشی توسعه و ترویج علوم کشاورزی، منابع طبیعی و محیط زیست ایران, تهران. https://civilica.com/doc/565203

Milk Quality Prediction Using CatBoost and Genetic Algorithm

Amir Hosein Esmaielpour

PhD Student, Department of Industrial Engineering, Faculty of Engineering, Kharazmi University, Tehran, Iran

Mariam Ameli

Assistant Professor of Industrial Engineering, Department of Industrial Engineering, Faculty of Engineering, Kharazmi University, Tehran, Iran

Abstract

Milk is one of the most important goods in the food chain in the household basket. Milk quality is a critical factor in ensuring consumer health and enhancing the performance of the dairy industry. In this study, the CatBoost algorithm was employed as a machine learning model to classify milk quality into three categories: “High,” “Medium,” and “Low.” To improve the model’s performance, hyperparameter tuning was performed using a genetic algorithm, which significantly increased the prediction accuracy. Furthermore, SHAP was utilized to enhance model interpretability, and its analysis revealed that the “Fat” feature is the most influential factor affecting the model’s output. The model achieved excellent performance with average evaluation metrics as follows: Precision of ۰٫۹۸٦۷, Recall of ۰٫۹۹۱۵, F۱-Score of ۰٫۹۸۸۹, and Accuracy of ۰٫۹۹۰٦. These results indicate the high accuracy and reliability of the proposed approach, suggesting that it can serve as a fast, cost-effective, and dependable tool for improving quality control processes in the dairy industry.

Keywords: milk quality prediction, machine learning, genetic algorithm, quality control