Kafkas Universitesi Veteriner Fakultesi Dergisi
Journal Home-Page: http://vetdergikafkas.org
E-ISSN: 1309-2251

Kafkas Univ Vet Fak Derg
32 (5): 727-738, 2026
DOI: 10.9775/kvfd.2026.37347

Research Article

Estimating Goat Milk Production in Türkiye: A Comparative Analysis of Time Series and Machine Learning Approaches for Accurate Forecasting

Köksal KARADAŞ [1], Osman Doğan BULUT [1 (*)], Hakan DUMAN [2], Selim TAKUR [3], Ayşe KARADAĞ GÜRSOY [1], Zeynep ÇELİK KAYSİM [1], Buse Eda AKYÜZ ULUCAN [4]

[1] Iğdır University, Faculty of Agriculture, Agricultural Economics Department, TR-76000 Iğdır - TÜRKİYE
[2] Iğdır University, Iğdır Vocational School, Transportation Services Department, TR-76000 Iğdır - TÜRKİYE
[3] Iğdır University, Iğdır Vocational School, Finance-Banking and Insurance Department, TR-76000 Iğdır - TÜRKİYE
[4] Akdeniz University, Faculty of Economics and Administrative Sciences, Econometrics Department, TR-07058 Antalya - TÜRKİYE

(*) **Corresponding author:**
Osman Doğan Bulut
Phone: +90 476 223 00 10 (4041)
Cellular phone: +90 542 552 54 29
E-mail: osman.dogan.bulut@igdir.edu.tr

How to cite this article?

Karadaş K, Bulut OD, Duman H, Takur S, Karadağ Gürsoy A, Çelik Kaysim Z, Akyüz Ulucan BE: Estimating Goat Milk Production in Türkiye: A Comparative Analysis of Time Series and Machine Learning Approaches for Accurate Forecasting. *Kafkas Univ Vet Fak Derg*, 32 (5): 727-738, 2026.
DOI: 10.9775/kvfd.2026.37347

Article ID: KVFD-2026-37347
Received: 18.06.2026
Accepted: 08.10.2026
Published Online: 21.10.2026

Abstract

Goat milk is a valuable food source because of its digestibility, nutritional value and suitability for value-added dairy products. This study forecast Türkiye's annual raw goat milk production for 2025-2029 and compared ARIMA with Decision Tree, MARS, XGBoost and SVM models using FAOSTAT data for 1961-2024. Nested expanding-window cross-validation, in which hyperparameters were tuned within each training window without access to the test year, produced 36 one-year-ahead forecasts for 1989-2024. Although MARS had the lowest errors (RMSE 33.07 versus 34.73 thousand tonnes for ARIMA), no machine-learning model differed significantly from ARIMA in Holm-adjusted Diebold–Mariano tests. The parsimonious ARIMA (2,1,0) model was therefore retained for probabilistic projections. Its point forecast declined from 468.22 thousand tonnes in 2025 to 420.39 thousand tonnes in 2029, whereas the 95% residual-bootstrap interval widened to 204.46-608.23 thousand tonnes, spanning nearly the entire historical range of the series. Sensitivity checks using post-2009 data and a fixed rolling window reproduced these conclusions. The forecasts therefore indicate substantial uncertainty rather than a firm decline and do not by themselves support specific policy measures; productivity, value-added processing and agricultural data systems warrant further investigation.

Keywords: Agricultural planning, ARIMA, Cross-validation, Forecasting, Goat milk, Machine learning, Model comparison, Time series analysis, Türkiye

Introduction

Goat milk is valued for its easy digestibility and for the production of high-value cheeses [1,2]. Its nutritional significance has three aspects: it is an important food source in developing countries affected by hunger and malnutrition, an alternative for consumers with allergies or gastrointestinal disorders in developed countries, and a product sought by health-conscious consumers [3]. Goat milk has been reported to support physiological functions in children and the elderly and to be consumed by individuals with milk allergies without adverse symptoms [4], and studies have addressed its nutritional properties, health benefits, production methods and comparison with other milk types [5].

India, Sudan, Pakistan, Bangladesh and France are the leading producers of raw goat milk, and Türkiye ranks seventh, with 482.25 thousand tonnes in 2024 [6]. In these countries goat farming is closely linked to rural livelihoods. In India, indigenous dual-purpose breeds support small and marginal farmers [7,8], and goat milk is popularly believed, without scientific substantiation, to help in the treatment of dengue fever [9]. Milk consumption is embedded in traditional dietary and pastoral practices in Sudan [10]; rising demand and the contribution of goat farming to household income and employment have increased interest in the activity in Pakistan [11,12]; and goat farming is economically important in Bangladesh despite limited production technology and the socio-economic constraints of farmers [13,14]. France has the most organized goat milk market in Europe [15]; its sector specializes in largely industrial cheese production, and milk accounts for 90-93% of gross income on dairy goat farms [16,17].

This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

In Türkiye, goat husbandry has a long history on the Anatolian Plateau, reflecting the country's geography and rural socio-economic structure, and goat products are a major source of animal protein in mountainous regions [18]. Forecasting studies of milk production have relied mainly on time-series models. ARIMA was selected for total milk production in India [19], a hybrid ARIMA-ANN model was the most accurate for buffalo milk production [20], autoregressive models predicted weekly goat milk yield at farm level [21], and ARIMA was more accurate than artificial neural networks (ANN) for goat and total milk production in Karnataka [22]. Goat milk yield and composition also vary across lactation [23], and trend analyses in Türkiye have reported upward trends in several animal products [24]. Machine-learning applications in the goat-milk sector, by contrast, have mostly addressed antioxidant peptides [25], adulteration detection [26], bulk-tank milk quality [27] and adaptive responses of goats to environmental conditions [28] rather than production forecasting.

Two gaps therefore remain. Empirically, the earlier national-level forecasts of Kaygısız and Sezgin [29] relied on data ending in 2016 and could not incorporate the 2021 peak and the subsequent decline. Methodologically, although algorithms such as XGBoost have been applied to Turkish agricultural data [30,31], comparative forecasting studies have generally ranked models by point-error measures such as RMSE [32,33]; such rankings do not show whether differences in accuracy exceed sampling variability, which can be assessed formally with the Diebold-Mariano test [34,35].

The scientific question of this study was whether flexible machine-learning algorithms provide a statistically detectable forecasting advantage over a parsimonious linear benchmark in a short annual production series. The null hypothesis of equal one-step-ahead predictive accuracy between each machine-learning model and ARIMA was tested. The aims were to compare four machine-learning techniques with ARIMA using nested rolling-origin cross-validation on updated FAOSTAT data (1961-2024) and to forecast Türkiye's annual raw goat milk production for 2025-2029 with quantified uncertainty.

Material and Methods

Ethical Approval

Ethics committee approval was not required for this study.

Data Source and Study Design

Annual raw goat milk production data for Türkiye (1961-2024, tonnes) were obtained from the FAOSTAT database [6].

Four machine-learning algorithms (Decision Tree, Multivariate Adaptive Regression Splines, Support Vector Regression and Extreme Gradient Boosting) were compared with an ARIMA benchmark using nested expanding-window cross-validation, and the selected specification was used for probabilistic forecasts for 2025-2029.

Data Preparation and Time-Series Diagnostics

Stationarity of the untransformed series was assessed with the Augmented Dickey-Fuller (ADF), Phillips-Perron (PP) and KPSS tests [36-38]; ADF and PP test a unit-root null hypothesis and KPSS a stationarity null. Because structural change can reduce the power of conventional tests, the Zivot-Andrews one-break test and the Lee-Strazicich minimum-LM tests with one and two endogenous breaks were also applied [39-41]. Break tests were applied to the level series, and conventional tests to the level and first difference. The break tests were used diagnostically only: the candidate break year was not used to restrict the cross-validation training windows, the machine-learning features or the final ARIMA estimation, all of which used the complete 1961-2024 series.

For the machine-learning models, the response was the annual first difference in production (Δyt) and the predictors were its first and second lags ($\Delta yt-1$ and $\Delta yt-2$), so every feature was observable at the forecast origin. Full-sample OLS residuals were not used as predictors. For MARS and SVM, the Yeo-Johnson transformation [42] and z-score normalization was applied to the two predictors to reduce skewness and place them on a comparable scale, and were fitted only on the data in each training sample; the tree-based models used the untransformed predictors. Because the response was not transformed, no back-transformation was required when forecasts were converted to production levels.

Forecasting Models

General background on the machine-learning methods and on resampling-based control of overfitting is given in [43].

Decision Tree Algorithms (DT): Decision trees partition the predictor space into homogeneous regions and can be used for classification and regression [44-46]; the CART algorithm was used [47].

Multivariate Adaptive Regression Splines (MARS): MARS represents nonlinear relationships by adding hinge functions and then removing terms that do not improve the fit [43,48]. Because of the small number of annual observations, prod_degree was fixed at 1 (main effects only) in the primary model; a sensitivity run allowed prod_degree = 2.

XGBoost Algorithm (XGB): XGBoost builds an ensemble of decision trees sequentially, each new tree being fitted to the errors of the current ensemble [49,50].

Support Vector Machines (SVM): Support Vector Regression with a radial basis function (RBF) kernel was used [51-53];

it fits a function whose deviations from the observed values do not exceed a tolerance (ε) while penalizing model complexity.

ARIMA (AutoRegressive Integrated Moving Average): ARIMA was used as the univariate benchmark [19,54].

Hyperparameter Optimization and Nested Expanding-Window Validation

A nested expanding-window procedure separated hyperparameter tuning from model evaluation. After lag construction, the data set contained 61 observations (1964-2024). The first outer training window included 25 observations, and each model was tested on the following year, giving 36 one-step-ahead forecasts for 1989-2024. Within each outer training sample, an inner expanding-window resampling initially using 65% of the training observations (minimum inner initial window 12 observations; at least 5 inner assessments) selected hyperparameters without access to the outer assessment year. Up to 30 randomly drawn candidates were evaluated for the general machine-learning grids, and 14 candidates for the primary MARS model. Preprocessing parameters were learned inside each resample. ARIMA was re-estimated at each outer forecast origin; d was chosen by the sequential KPSS rule (maximum d = 1), and p and q by an exhaustive AICc search over non-seasonal models. Model comparisons used only the pooled predictions from the outer test years [55-57]. Validation was based on one-step-ahead (h = 1) forecasts, whereas the final projections covered h = 1-5.

To limit overfitting in this small sample [43], the hyperparameter search ranges were restricted and kept identical in every training window (maximum tree depth 5 for the Decision Tree and 4 for XGBoost), interaction terms were excluded from the primary MARS model, the predictor set was limited to two lagged differences, and hyperparameter selection was separated from performance evaluation by the nested design.

For each tunable model, candidates were drawn from the search ranges in *Table 1* using random grids (values drawn from uniform or log-uniform distributions, oversampled and de-duplicated, then randomly subsampled to the target grid size); the small discrete MARS grid was evaluated exhaustively. Within each outer training sample, the configuration with the lowest pooled inner RMSE (computed on the differenced scale; ties broken by pooled inner MAE) was selected. *Table 1* reports the search range for each hyperparameter and the value selected when each model was refit to the full 1964-2024 sample.

Structural-Break Sensitivity Checks

Two checks assessed the diagnostic-only treatment of the candidate 2009 break. First, ARIMA was re-estimated using only post-break observations (2009-2024; 16 observations), and its order and 2025-2029 forecasts were compared with those of the full-sample model. A single refit was used because lag construction leaves only 13 post-break observations, too few for nested cross-validation; its intervals were obtained with the same 5.000-path i.i.d. residual bootstrap, without the residual-diagnostic selection rule. Second, the complete model comparison, with identical hyperparameter search, was repeated using a fixed 25-observation rolling training window for the machine-learning models, producing 36 one-step-ahead forecasts (1989-2024) and the same Holm-adjusted Diebold-Mariano tests. ARIMA was re-estimated from all observations available before each forecast origin under both schemes, so its errors were identical in the two analyses.

Forecast Accuracy Metrics and Statistical Comparison

Forecast accuracy was measured with RMSE, MAE, MAPE and bias on the original production scale. Forecasts of annual changes were converted to production levels by adding them to the last observed value at each forecast origin. All measures were computed from the 36 pooled outer-fold errors, because each outer test fold contained a single observation [33,55].

Error Metrics

- Forecast Error: the difference between the observed and predicted production level at horizon i from origin T. Written as: $\varepsilon(T+i) = y(T+i) - \hat{y}(T+i)$.

Table 1. *Hyperparameter search ranges and final selected values for the four tunable machine-learning models*

Model	Hyperparameter(s) and Search Range	Final Selected Value(s), Full-Sample Refit
Decision Tree	cost_complexity: 10^U(−6,−1); tree_depth: {1..5}; min_n: {2..12}	cp=3.94e-05; depth=2; min_n=6
MARS	num_terms: {2..15} (14 candidates); prod_degree: fixed at 1 for the primary model (1:2 evaluated only in the separate sensitivity run)	num_terms=3; prod_degree=1
SVM (RBF)	cost: 2^U(−5,7); rbf_sigma: 10^U(−5,0); margin (ε): 10^U(−3,$\log_{10}$ 0.5)	cost=65.661; sigma=0.00829; margin=0.347
XGBoost	trees: {10,20,...,500}; tree_depth: {1..4}; learn_rate: 10^U(−3,$\log_{10}$ 0.3); loss_reduction (γ): 10^U(−6,1); min_n: {2..10}; sample_size: U(0.60,1.00)	trees=340; depth=2; min_n=6; rate=0.0036; gamma=0.0577; sample=0.946

Search ranges follow standard notation U(a,b) for a uniform draw on [a,b] and 10^U(a,b)/2^U(a,b) for a log-uniform draw exponentiated on base 10 or 2, respectively; braces {.} denote a discrete uniform draw over the listed integer set

- RMSE (Root Mean Squared Error): $RMSE = \sqrt{[(1/n) \times \Sigma i\ \varepsilon(T+i)^2]}$, with n = 36 pooled outer-fold predictions.
- MAE (Mean Absolute Error): $MAE = (1/n) \times \Sigma i\ |\varepsilon(T+i)|$
- MAPE (Mean Absolute Percentage Error): $MAPE = 100 \times (1/n) \times \Sigma i\ |\varepsilon(T+i)/y(T+i)|$, where $y(T+i)$ is the observed production level [55].

Pairwise forecast accuracy was assessed using the Harvey-Leybourne-Newbold small-sample modification of the Diebold-Mariano test with a one-step horizon (h = 1) and squared-error loss [34,35]. The tests used the aligned production-level errors from the 36 outer folds. Because ARIMA was compared with four alternative models, Holm-adjusted P-values were additionally reported to control family-wise error. As a robustness check, the tests were repeated under absolute-error loss.

Final Model Estimation and Forecast Uncertainty

After model comparison, ARIMA was re-estimated on all observations up to 2024. Residual autocorrelation was assessed with a Ljung-Box test (fitdf-corrected for the estimated AR and MA orders) and conditional heteroscedasticity with a Ljung-Box test on the squared residuals (McLeod-Li test; 10 lags). Under a pre-specified rule, 95% forecast intervals were obtained from a 5.000-path i.i.d. residual bootstrap if neither test was significant at the 5% level, and from the analytic Gaussian approximation otherwise. Residual normality was examined with the Shapiro-Wilk and Jarque-Bera tests, and analytic Gaussian intervals were computed for comparison *(Table 3)*.

The four machine-learning models were also refit to the complete 1964-2024 sample with the hyperparameters in *Table 1*; only ARIMA was used for the 2025-2029 forecasts.

Software

Analyses were conducted in R 4.5.3 under Debian GNU/Linux 13. The principal package versions were tidyverse 2.0.0, tidymodels 1.4.1, forecast 9.0.2, earth 5.3.5, kernlab 0.9-33, xgboost 3.4.2.1, rpart 4.1.27, urca 1.3-4, and future 1.70.0. The structural-break sensitivity checks were run in a separate script in the same R session.

Results

Model Evaluation and Selection

Türkiye's position among the world's raw goat milk producers in 2024 is shown in *Fig. 1*. Raw goat milk production in Türkiye declined in the early 1960s, remained relatively stable from the mid-1960s to the late 1970s and decreased from the 1980s to its lowest value in 2009 (192.21 thousand tonnes). It then increased until 2021 (622.79 thousand tonnes) and has declined since *(Fig. 2-A)*.

The conventional tests indicated that the series was integrated of order one: ADF and PP did not reject a unit root at the level but rejected it for the first difference, whereas KPSS rejected level stationarity but not first-

Table 2. *Pooled outer rolling-origin forecast performance and selected final specifications*

Model	Selected Final Specification	RMSE	MAE	MAPE (%)	Bias
MARS	num_terms=3; prod_degree=1	33.07	21.25	6.05	4.53
SVM	cost=65.661; sigma=0.00829; margin=0.347	34.11	22.51	6.37	4.55
XGBoost	trees=340; depth=2; min_n=6; rate=0.0036; gamma=0.0577; sample=0.946	34.36	22.76	6.44	6.67
ARIMA	ARIMA (2,1,0) final refit	34.73	22.88	6.61	4.75
Decision Tree	cp=3.94e-05; depth=2; min_n=6	36.92	24.84	7.11	4.05

RMSE, MAE, and bias are given in thousand tonnes; MAPE is given as a percentage. The figures are based on 36 one-year-ahead forecasts covering 1989-2024. Errors were calculated after the forecasts had been converted back to production levels. The specifications in the table come from models refitted to the full sample and therefore may differ from those selected in individual folds. Bias was calculated as actual production minus forecast production. A positive value means that production was underestimated on average

Table 3. *Analytic Gaussian versus i.i.d. residual-bootstrap 95% interval widths for the final ARIMA (2,1,0) forecasts*

Year	Analytic Gaussian 95% Width	i.i.d. Residual-bootstrap 95% Width
2025	110.51	165.53
2026	176.42	218.79
2027	251.41	287.70
2028	319.93	346.91
2029	385.80	403.77

Interval widths (Upper 95 minus Lower 95) are in thousand tonnes, computed from the analytic Gaussian and i.i.d. residual-bootstrap forecasts produced for the same final ARIMA (2,1,0) specification

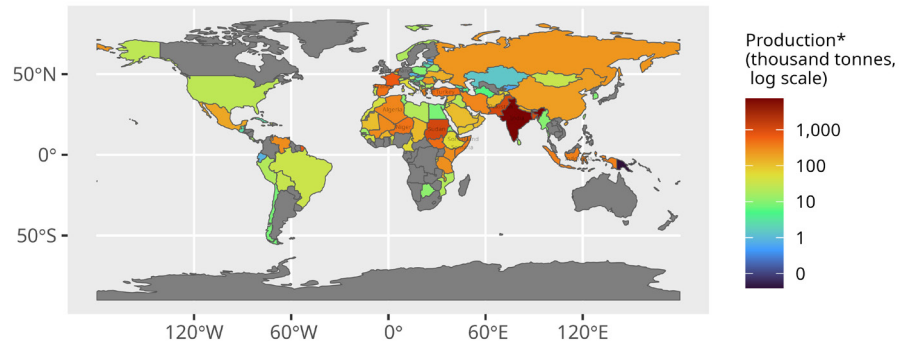

Fig 1. World raw goat milk production in 2024. Map produced by the authors from FAOSTAT data [6], available under CC BY 4.0; thousand tonnes; colours are shown on a logarithmic scale

difference stationarity *(Table 4)*. Zivot-Andrews rejected the unit root only in the intercept-break model, with a break in 2009; the trend-break and combined-break models and all Lee-Strazicich specifications did not reject it *(Table 4)*. First differencing was therefore retained for the primary models.

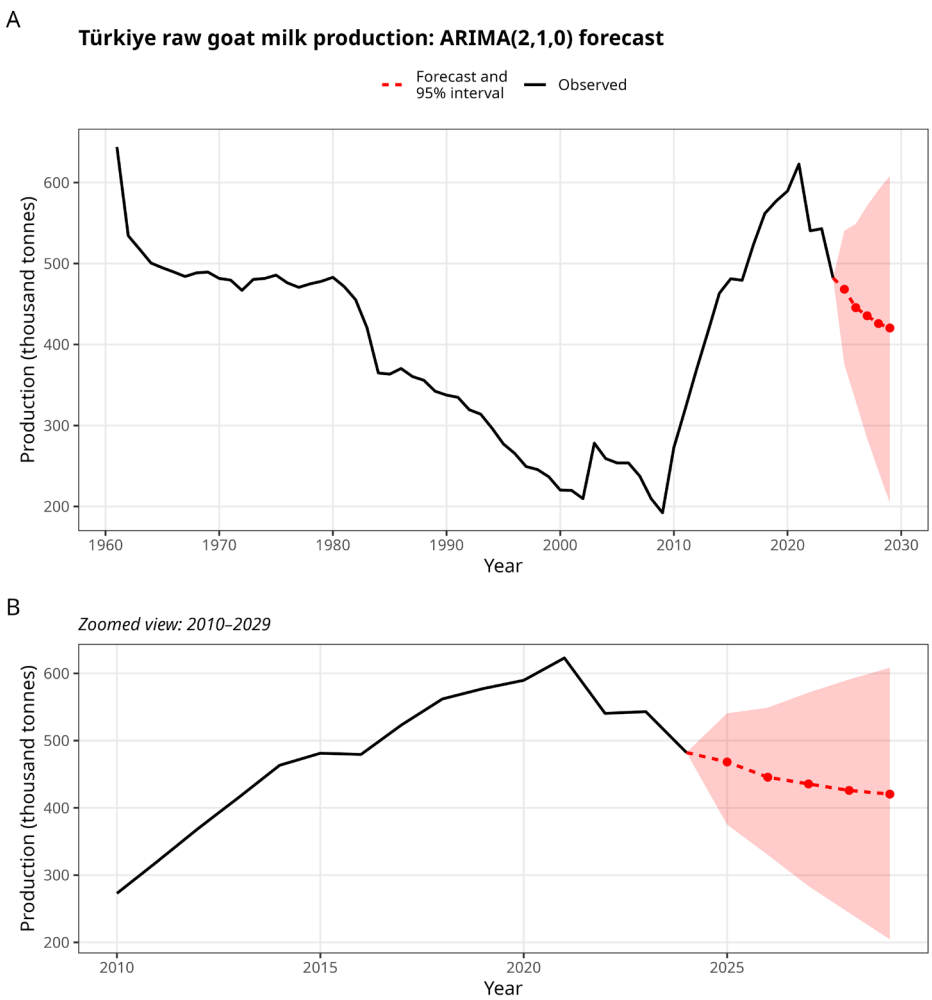

Fig 2. Observed and forecast raw goat milk production in Türkiye. **(A)** Full time series with observed values (solid black line) and ARIMA (2,1,0) forecasts for 2025-2029 (red dashed line and points), together with the diagnostically selected 95% residual-bootstrap interval (shaded region) based on 5.000 simulated paths; **(B)** Zoomed view of the 2010-2029 period. Values are expressed in thousand tonnes. Figure produced by the authors from FAOSTAT data [6] (CC BY 4.0) and the forecasts of this study

Pooled performance of the 36 one-step-ahead outer forecasts (1989-2024) is given in *Table 2*.

MARS had the lowest RMSE, MAE and MAPE, with an RMSE 4.79% lower than that of ARIMA; SVM, XGBoost and ARIMA had similar errors, and the Decision Tree had the highest *(Table 2)*. None of the four Diebold-Mariano comparisons with ARIMA was significant, before or after Holm adjustment *(Table 5)*, and the conclusion was unchanged under absolute-error loss (DM statistics −0.983 to 0.941; raw P≥0.332; all Holm-adjusted P=1.000). All models had a positive mean bias *(Table 2)*. For every model, the two largest absolute errors occurred in 2010 and 2022, the years following the 2009 trough and the 2021 peak: production was underpredicted in 2010 (by 87.67-98.51 thousand tonnes) and overpredicted in 2022 (by 89.63-112.75 thousand tonnes). The largest single error accounted for 20.1-25.9% of each model's pooled squared error.

The unrestricted MARS variant (prod_degree = 1:2) gave the same predictions as the primary model in 33 of 36 folds, but its pooled RMSE was more than twice as high *(Table 6)*. This difference arose mainly from the 2012 test year, in which the lagged-difference input fell outside the training range: the unrestricted variant predicted −19.01 thousand tonnes, whereas the primary model predicted 316.48 thousand tonnes (actual: 369.43 thousand tonnes).

Table 4. *Unit-root and structural-break test results for the annual goat milk production series, 1961-2024*

Test	Series and Specification	Statistic	5% Critical Value	Break Year(s)	Decision at 5%
ADF	Level; constant and trend	−1.750	−3.450	–	Unit root not rejected
PP	Level; constant and trend	−1.706	−3.481	–	Unit root not rejected
KPSS	Level; constant and trend	0.156	0.146	–	Stationarity rejected
ADF	First difference; constant	−3.059	−2.890	–	Unit root rejected
PP	First difference; constant	−6.769	−2.908	–	Unit root rejected
KPSS	First difference; constant	0.288	0.463	–	Stationarity not rejected
Zivot-Andrews	Level; intercept break	−5.918	−4.800	2009	Unit root rejected
Zivot-Andrews	Level; trend break	−3.636	−4.420	2004	Unit root not rejected
Zivot-Andrews	Level; intercept and trend break	−3.591	−5.080	2003	Unit root not rejected
LS, one break	Level; Model A (level break)	−1.807	−3.566	2003	Unit root not rejected
LS, one break	Level; Model C (level and trend break)	−3.812	−4.463	2010	Unit root not rejected
LS, two breaks	Level; Model A (level breaks)	−2.023	−3.842	2003; 2010	Unit root not rejected
LS, two breaks	Level; Model C (level and trend breaks)	−4.914	−5.730	2006; 2016	Unit root not rejected

ADF, augmented Dickey-Fuller (lag length selected by BIC, maximum 4); PP, Phillips-Perron and KPSS (bandwidth 10); LS, Lee-Strazicich minimum LM test (lag selection general-to-specific, maximum 4). ADF, PP, Zivot-Andrews, and LS test the null of a unit root; KPSS tests the null of stationarity. The break tests were used diagnostically only (see Materials and Methods)

Table 5. *Primary (expanding-window) Diebold–Mariano tests against the ARIMA benchmark, all four comparisons*

Comparison (ARIMA vs.)	DM Statistic	Raw P-value	Holm-adjusted P
MARS	0.943	0.352	1.000
SVM	0.299	0.767	1.000
XGBoost	0.176	0.861	1.000
Decision Tree	−0.947	0.350	1.000

Figures are taken directly from the pooled outer-fold predictions of the primary analysis

Table 6. *MARS interaction-term (prod_degree) sensitivity check*

Measure	MARS (prod_degree=1, primary)	MARS (prod_degree=1:2, unrestricted)
Pooled outer-fold RMSE (thousand tonnes)	33.07	72.14
Folds with a different prediction than the other variant (of 36)	–	3
Worst single-fold outcome (2012 test year)	316.48 (valid; error +52.95)	−19.01 (invalid: negative production)

RMSE and forecast values are in thousand tonnes

The difference between the two variants was not significant (DM = −0.999, raw P=0.324, Holm P=0.324).

Re-estimation on the post-break observations (2009-2024) selected the same ARIMA (2,1,0) order. Its forecasts were close to the full-sample forecasts in 2025 but lower thereafter (376.19 versus 420.39 thousand tonnes in 2029), and its interval was wider at the five-year horizon *(Table 7)*.

With a fixed 25-observation rolling window, the pooled RMSE and MAPE of all four machine-learning models were higher than with the expanding window; the increase was largest for the Decision Tree and SVM and smallest for XGBoost and MARS *(Table 8)*. MARS again had the lowest pooled error, and none of the Holm-adjusted Diebold-Mariano comparisons with ARIMA was significant *(Table 9)*.

Final Production Forecasts

The full-sample ARIMA (2,1,0) model (AIC = 1474.20, AICc = 1474.61, BIC = 1480.63) showed no residual autocorrelation (Ljung-Box Q* = 6.517, df = 8, P=0.590) and no squared-residual dependence (McLeod-Li, 10 lags, statistic = 2.637, P=0.989); residual-bootstrap intervals were therefore used. Residual normality was rejected (Shapiro-Wilk W = 0.871, P<0.001; Jarque-Bera = 54.50, df = 2, P<0.001), mainly because of excess kurtosis (7.48; skewness −0.31), and three of the 64 standardized residuals exceeded 3 in absolute value. The bootstrap intervals were wider than the analytic Gaussian intervals at every horizon *(Table 3)*.

The ARIMA (2,1,0) point forecast declined from 468.22 thousand tonnes in 2025 to 420.39 thousand tonnes in 2029, while the 95% residual-bootstrap interval widened from 375.08-540.60 to 204.46-608.23 thousand tonnes *(Table 7; Fig. 2)*. The 2029 interval includes values above the 2024 production level (482.25 thousand tonnes) and spans most of the range observed in the historical series *(Fig. 2-A)*.

Discussion

MARS ranked first on all three point-error measures, but its RMSE advantage over ARIMA (4.79%) was not statistically significant, and none of the four machine-learning models differed significantly from ARIMA

Table 7. *Full-sample versus post-break-only ARIMA (2,1,0) forecasts, 2025-2029*

Year	Full-sample ARIMA (2,1,0) Forecast [95% CI]	Post-break-only ARIMA (2,1,0) Forecast [95% CI]
2025	468.22 [375.08–540.60]	469.99 [375.52–528.11]
2026	445.53 [330.08–548.87]	432.10 [304.95–527.43]
2027	435.54 [283.64–571.34]	416.43 [226.80–556.53]
2028	425.91 [243.94–590.85]	391.00 [149.95–577.54]
2029	420.39 [204.46–608.23]	376.19 [70.47–616.03]

Figures are taken directly from the structural-break sensitivity checks described in Materials and Methods; forecast values are in thousand tonnes

Table 8. *Pooled outer-fold performance under the expanding-window (primary) and fixed rolling-window training schemes*

Model	Expanding RMSE	Rolling RMSE	Expanding MAPE (%)	Rolling MAPE (%)
MARS	33.07	34.62	6.05	6.26
ARIMA	34.73	34.73*	6.61	6.61*
XGBoost	34.36	34.82	6.44	6.49
SVM	34.11	38.48	6.37	7.05
Decision Tree	36.92	41.74	7.11	7.72

*RMSE is in thousand tonnes and MAPE in percent. *ARIMA's rolling-window and expanding-window metrics are identical by construction (see Methods). Expanding-window values are those reported in Table 2*

Table 9. *Rolling-window Diebold-Mariano tests against the ARIMA benchmark*

Comparison (ARIMA vs.)	DM Statistic	Raw P-value	Holm-adjusted P
MARS	0.048	0.962	1.000
XGBoost	−0.046	0.963	1.000
SVM	−1.430	0.161	0.646
Decision Tree	−1.128	0.267	0.800

Figures are taken directly from the rolling-window structural-break sensitivity check described in Materials and Methods

after Holm adjustment. These results should be read as a ranking without statistical evidence of a difference, not as evidence of equivalence: a non-significant Diebold–Mariano test indicates only that the 36 pooled one-step-ahead errors did not provide sufficient evidence of a difference, and with so few errors the test has limited power to detect differences of the size observed, even with the modified Diebold-Mariano statistic recommended by Harvey et al.[35]. ARIMA was therefore retained for the final projections neither because the test favoured it nor because it ranked first, but for reasons that do not depend on the test result: it is the most parsimonious specification, with directly interpretable parameters; it yields coherent multi-step prediction intervals directly [55], whereas the machine-learning models were set up here only as one-step-ahead forecasters of annual changes and do not provide comparable probabilistic multi-step forecasts; and its point-forecast accuracy was close to that of the best-ranked model. The choice is therefore a practical one. It does not imply that ARIMA was the most accurate point forecaster, that all models performed equally, or that differencing removed all nonlinear behaviour or structural change from the series; a direct comparison of the five-year trajectories implied by MARS and ARIMA was not possible within the present one-step framework and is noted as a limitation below.

ARIMA has also performed well in earlier forecasts of aggregate milk production [19], and Kumar et al. [22] reported ARIMA to be more accurate than ANN for goat and total milk production, whereas an ARIMA-ANN hybrid was the most accurate model for buffalo milk production in India [20]. In Turkish livestock series, ARIMA also provided the most successful forecasts of goat red-meat production in a comparison with exponential smoothing and STLF methods [32]. At a much finer temporal scale, autoregressive models based on current and previous yields predicted weekly goat milk yield well in a farm-level study [21], although that setting is not directly comparable to annual national aggregates. In our analysis, no model was clearly superior, and the machine-learning models were retained as out-of-sample comparators rather than used to generate the 2025-2029 projections. Evidence from other domains is mixed: in seasonal rainfall forecasting in India, XGBoost was more accurate than ARIMA and several other machine-learning models [50], whereas in an early comparison of support vector machines, recurrent neural networks and ARIMA models, ARIMA performed best on linearly generated data, support vector machines gained an advantage only on a strongly nonlinear chaotic benchmark, and the two performed equally on a real industrial series [53]. The absence of a detectable advantage for the machine-learning models here is consistent with, but does not demonstrate, a largely linear dynamic in the differenced production series. Updating the series to 2024 matters because the 2021 peak and the subsequent decline are now part of the observed data: the ARIMA forecast for 2026 (445.53 thousand tonnes) is lower than the 495.54 thousand tonnes projected by Kaygısız and Sezgin [29] from data ending in 2016. Because the current 95% interval for 2026 (330.08-548.87 thousand tonnes) includes their estimate, this difference is not sufficient to conclude that the earlier forecast was inaccurate. Using breed-level Turkish Statistical Institute data for 1991-2019, Serin and Siren [24] found that trend tests disagreed on the direction of Anatolian Black goat milk production, with some indicating an upward and others a downward trend; because their series ended in 2019 and differed in source and coverage from the national FAOSTAT totals used here, it did not include the 2021 peak and subsequent decline, and the two analyses are not directly comparable. Finally, Şahin Demirel and Erik [31] used XGBoost to examine the effect of climatic variables on small-ruminant milk yield at the provincial level; multivariate applications of this kind complement the univariate national-level comparison presented here.

Goat milk production rose after the 2009 trough, peaked in 2021, and has declined since then [6]. Goat milk supply in Türkiye rests largely on the indigenous Hair goat, which accounts for most of the national herd, whereas specialised dairy genotypes such as Maltız and Saanen are limited in number and concentrated in the Aegean and Marmara regions [18]. Structural factors of this kind are not observed by a univariate model, and their contribution to the turning points in the series could not be assessed here. The ARIMA point forecasts continue this decline gradually, from 468.22 thousand tonnes in 2025 to 420.39 thousand tonnes in 2029, but the residual-bootstrap intervals widen considerably as expected for a differenced (d = 1) ARIMA model whose prediction intervals continue to grow with the forecast horizon [55], and, by 2029, are compatible with further decline, stabilization, or recovery; the point forecasts are therefore best read as one plausible path rather than a firm prediction. Because the final-model residuals were heavy-tailed, the residual-bootstrap intervals, which do not assume normality, were wider than the analytic Gaussian intervals at every horizon *(Table 3)*, indicating that a Gaussian approximation would have understated forecast uncertainty. All models also had a positive average bias, that is, they underpredicted production on average *(Table 2)*. Because every model forecasts annual changes from the most recently observed level, it can respond to a turning point only after that turning point has been observed; consistent with this, the two largest errors of every model occurred in 2010 and 2022, immediately after the 2009 trough and the 2021 peak.

Methodologically, the study combines nested rolling-origin validation, which separates hyperparameter tuning from performance evaluation [44] and follows recommendations to use rolling origins, recalibration at each origin and multiple test periods when evaluating forecasts on a single series [56,57], with Holm-adjusted Diebold-Mariano tests, which assess whether ranking differences exceed what would be expected from sampling variability [34]. Earlier comparative forecasting studies of Turkish livestock and crop series identified the best-performing method for each series [32] or ranked models by test-set RMSE from a single training–testing split [33]; the present results show that a first-place ranking need not be statistically distinguishable from a parsimonious benchmark. While machine-learning algorithms continue to be applied in regional yield modelling and climate-livestock studies [31], and, in the goat sector, mostly to tasks other than production forecasting, such as adulteration detection [26], milk-quality prediction [27] and modelling adaptive responses to climatic conditions [28], their added value for short univariate macro-level series should be demonstrated empirically rather than assumed.

The MARS interaction-term sensitivity check provides empirical support for the a priori decision to fix prod_degree at 1 in the primary MARS specification. Allowing interaction terms changed the prediction in only 3 of 36 folds, so the practical cost of this restriction was minimal in most years; however, because MARS represents interactions as products of spline basis functions [48], an input outside the training range enters the prediction through a product term; in the one fold where the lagged input fell outside the range observed in training, permitting interactions led the model to extrapolate to a physically impossible negative production forecast, more than doubling the pooled RMSE. This single fold dominated the mean-squared-error comparison of the two variants, yet the Diebold-Mariano test did not identify the difference as significant. It also supports preferring the simpler, more stable specification when the training sample is this small [44] The Decision Tree ranked last on all three point-error measures; because a regression tree predicts a constant value within each terminal node [44], the selected depth-2 tree could return at most four distinct predicted changes, limiting its ability to follow the series.

The Zivot-Andrews test, which estimates the break date from the data [39], indicated a candidate intercept break around 2009, whereas the Lee-Strazicich tests, which allow for breaks under both the null and the alternative hypotheses so that a rejection unambiguously implies trend stationarity [40] and spurious rejections are avoided [41], did not reject the unit root under any specification (*Table 4*). Because this candidate break was used only diagnostically in the primary analysis rather than to restrict model estimation or validation, two sensitivity checks examined whether this choice mattered in practice. Re-estimating ARIMA from post-break data alone reproduced the same model order and a broadly similar, if somewhat lower and less certain, five-year forecast path, suggesting that the diagnostic-only treatment of the break did not conceal a materially different post-break dynamic. Repeating the full model comparison with a fixed rolling training window, rather than the expanding window used throughout [56], did not improve the machine-learning models' relative standing against ARIMA - if anything, every machine-learning model performed somewhat worse with less historical data to draw on, and none was significantly different from ARIMA under the same Holm-adjusted Diebold-Mariano procedure. Taken together, these checks support treating the reported model comparison and forecast trajectory as reasonably robust to the candidate structural break, while the wider intervals obtained under the post-break-only re-estimation reinforce, rather than contradict, the emphasis already placed on forecast uncertainty. These checks do not, however, model the break explicitly (for example, through break dummies or regime-specific parameters). The post-break analysis, based on a single refit, addresses the stability of the model order and forecast path rather than comparative accuracy, and the rolling-window check restricted the training history only for the machine-learning models, so it does not show whether a rolling-window ARIMA would perform differently.

A key limitation of this study is its univariate approach, which relies solely on historical production data. In addition, the machine-learning models were given only two lagged differences as predictors, a design similar to the three-lag autoregressive set-up used with log-differenced annual FAOSTAT data for cherry production in Türkiye [33]; this safeguards against overfitting with 61 observations, but it effectively reduces these algorithms to flexible second-order autoregressive functions and prevents them from exploiting their main strength, the combination of many predictors. With 61 observations after lag construction, flexible algorithms such as XGBoost and RBF-SVR also carry a risk of overfitting and of unstable, high-variance predictions [43]; the design restrictions described in the Methods limit but do not eliminate this risk, and with only 36 pooled errors small-sample instability cannot be excluded as an explanation of the non-significant Diebold-Mariano results. This design constraint, and not only the small sample, is a plausible explanation for the absence of significant differences from ARIMA. Future work could incorporate exogenous variables such as feed prices, climate conditions, and macroeconomic indicators, support policies and animal-health events, the latter two having been proposed as

factors behind trends in Turkish livestock production [24], using model classes designed for such inputs, for example ARIMAX- or SARIMAX-type dynamic regression models, structural and Bayesian structural time-series models, or panel-data models across provinces; in a multivariate comparison for raw milk supply in Türkiye, SARIMAX gave the most accurate forecasts, followed by XGBoost [30]. Hybrid approaches, such as modelling ARIMA residuals with a machine-learning model, analogous to the ARIMA-ANN hybrid that was most accurate for buffalo milk in India [20], were also not evaluated here.

A further limitation concerns the forecast horizon used for model evaluation: model comparison and selection, including the Diebold-Mariano tests reported above, were based exclusively on one-step-ahead (h=1) rolling-origin errors, whereas the final projections extend to a five-year horizon (h=1 through h=5). The relative ranking of ARIMA and the machine-learning alternatives at longer horizons was therefore not assessed within the same cross-validated framework, and the reported comparative accuracy should be read as evidence about one-year-ahead performance rather than about the full five-year forecast path; future work should evaluate multi-step accuracy directly. Finally, the sampling uncertainty of the error metrics themselves was not quantified (for example, through bootstrap confidence intervals for RMSE differences), although the Diebold-Mariano conclusions did not change when absolute-error loss was used instead of squared-error loss.

Because this study forecasts aggregate production without identifying the causal drivers of the projected trajectory, its results cannot determine which interventions would reverse or mitigate a decline. The width of the prediction intervals also argues for flexible, adaptive planning rather than commitments tied to a single point forecast. The following areas are therefore suggested for further policy investigation; they are consistent with, but not derived from, the forecasts, and their relevance does not depend on which forecast scenario materializes:

- Farm productivity and herd quality: whether support for small-ruminant farming that focuses on animal health, farm management, and productivity, including, where appropriate for local climate and production systems, the use of higher-yielding breeds (within Türkiye, the Kilis and Shami breeds are already recognised for higher milk yield than the dominant Hair goat [18]), can stabilize supply warrants dedicated evaluation.
- Value-added processing: if the supply of raw goat milk falls or remains uncertain, processing a larger share into higher-value products such as cheese, yoghurt, and ice cream [5] could raise income from the existing supply In European dairy goat sectors, milk is largely processed on-farm for national markets, with some protected-designation cheeses exported [17], and in some markets the commercial range of goat milk products, including fermented drinks and organic products, has been expanding [4]. Feasibility nevertheless requires separate market and cost analysis, particularly because marketing margins along the Turkish goat-product value chain have been described as irregular [18].
- Data infrastructure: annual aggregates [6] cannot capture within-year variation such as the lactation-stage changes in milk yield and composition reported for goats [23]; more frequent, granular, and multivariate production data would allow future studies to model the drivers of production directly and could improve the precision of forecasts.

In conclusion, none of the four machine-learning models provided a statistically detectable one-step-ahead accuracy advantage over a parsimonious ARIMA benchmark for Türkiye's annual goat milk production, and the ARIMA projections for 2025-2029 carry uncertainty wide enough to encompass further decline, stabilization, or recovery. Forecasts of this kind are most useful as a transparent, regularly updated reference for decision support, interpreted together with their uncertainty rather than as point targets.

Declarations

Availability of Data and Materials: The production data analyzed in this study are publicly available from the FAOSTAT database [6] under the Creative Commons Attribution 4.0 International license (CC BY 4.0); *Fig. 1* and *Fig. 2* were produced by the authors from these data. The analysis code (main analysis script and structural-break sensitivity module) and outputs are available from the corresponding author (ODB) on reasonable request.

Acknowledgements: The authors have no acknowledgements to declare.

Funding Support: No funding was received for this study.

Competing Interests: The authors declare that they have no competing financial or non-financial interests.

Declaration of Generative Artificial Intelligence (AI): We hereby declare that the main manuscript, including tables and figures, was not authored or created using AI or AI-assisted technologies. AI and AI-assisted technologies such as ChatGPT and DeepL were solely employed to enhance the readability and language of the article. The authors have carefully reviewed and edited all AI-assisted sections to ensure their reliability and accuracy. The authors take full responsibility for the content of the manuscript.

Ethical Compliance Declaration: The authors declare that this study involves the analysis of pre-existing, fully anonymized secondary data. No animal subjects were directly involved during this phase of the research, and all data were handled strictly in accordance with fundamental principles of data privacy and academic integrity.

Authors' Contributions: KK: Revision, editing, and finalization of the manuscript. All authors participated in the analysis process, ensuring the accuracy and validity of the results. ODB: Revision, editing, and finalization of the manuscript. HD, ST: Conceptualization, design, planning, investigation, data analysis, and writing the original draft. AKG, ZÇK: Revision, editing, and finalization of the manuscript. BEAU: Data analysis, revision and editing of the manuscript. They also reviewed and approved the final manuscript.

References

1. **Darwish HR, Darwish AM, Abdel-Salam AM, Lethy HM, El-Badawy MM, Zoheir KM:** PRL and CSN1S2 genes and milk traits in Damascus and Zaraibi goats. *Egypt J Vet Sci,* 54 (3): 465-470, 2023. DOI: 10.21608/EJVS.2023.192936.1442
2. **Liao G, Wang T, Li X, Gu J, Jia Q, Wang Z, Li H, Qian Y, Qiu J:** Comparison of the lipid composition of milk fat globules in goat *(Capra hircus)* milk during different lactations and human milk. *Foods,* 13 (11):1618, 2024. DOI: 10.3390/foods13111618
3. **Haenlein GFW:** Goat milk in human nutrition. *Small Rumin Res,* 51 (2): 155-163, 2004. DOI: 10.1016/j.smallrumres.2003.08.010
4. **Paszczyk B, Czarnowska-Kujawska M, Klepacka J, Tońska E:** Health-promoting ingredients in goat's milk and fermented goat's milk drinks. *Animals,* 13 (5):907, 2023. DOI: 10.3390/ani13050907
5. **Nayik GA, Jagdale YD, Gaikwad SA, Devkatte AN, Dar AH, Ansari MJ:** Nutritional profile, processing and potential products: A comparative review of goat milk. *Dairy,* 3 (3): 622-647, 2022. DOI: 10.3390/dairy3030044
6. **FAO:** Crops and livestock products. Food and Agriculture Organization of the United Nations, Rome. https://www.fao.org/faostat/en/#data/QCL; *Accessed:* 19.05.2026.
7. **Sato K:** Goat-rearing practices and the limited effects of the SHG program in India: evidence from a Tamil Nadu village. *Southeast Asian Stud,* 49 (1): 52-73, 2011. DOI: 10.20495/tak.49.1_52
8. **Verma AK, Rajkumar V:** Production, processing and marketing of goat milk in India: Reality and scope. *Indian J Small Rumin,* 27 (2): 152-159, 2021. DOI: 10.5958/0973-9718.2021.00031.3
9. **Pal M, Dudhrejiya TP, Pinto S, Brahamani D, Vijayageetha V, Reddy Y, Kate P:** Goat milk products and their significance. *Beverage Food World,* 44 (7): 21-25, 2017.
10. **Abdelgadir WS, Ahmed TK, Dirar HA:** The traditional fermented milk products of the Sudan. *Int J Food Microbiol,* 44 (1): 1-13, 1998. DOI: 10.1016/S0168-1605(98)00090-7
11. **Khan MS, Khan MA, Mahmood S:** Genetic resources and diversity in Pakistani goats. *Int J Agric Biol,* 10 (2): 227-231, 2008.
12. **Muhammad MS, Abdullah M, Javed K, Khan MS, Jabbar MA:** Goat production systems in Punjab, Pakistan. *J Anim Plant Sci,* 25 (3), 2015.
13. **Hossain M, Akhtar A, Hossain M, Choudhury M, Islam F:** Goat husbandry practices in Southern region of Bangladesh. *J Biosci Agric Res,* 5 (2): 59-64, 2015. DOI: 10.18801/jbar.050215.55
14. **Shoshe NZ, Islam MT, Mahmud MAA:** Socio economic condition of Black Bengal goat farmers and their goat management in rural areas of North Bengal region, Bangladesh. *Res Agric Livest Fish,* 6 (1): 111-118, 2019. DOI: 10.3329/ralf.v6i1.41391
15. **Miller BA, Lu CD:** Current status of global dairy goat production: An overview. *Asian-Australas J Anim Sci,* 32 (8): 1219-1232, 2019. DOI: 10.5713/ajas.19.0253
16. **Morgan F, Massouras T, Barbosa M, Roseiro L, Ravasco F, Kandarakis I, Bonnin V, Fistakoris M, Anifantakis E, Jaubert G:** Characteristics of goat milk collected from small and medium enterprises in Greece, Portugal and France. *Small Rumin Res,* 47 (1): 39-49, 2003. DOI: 10.1016/S0921-4488(02)00252-3
17. **Pulina G, Milán MJ, Lavín MP, Theodoridis A, Morin E, Capote J, Thomas DL, Francesconi AHD, Caja G:** Invited review: Current production trends, farm structures, and economics of the dairy sheep and goat sectors. *J Dairy Sci,* 101 (8): 6715-6729, 2018. DOI: 10.3168/jds.2017-14015
18. **Daşkıran I, Savas T, Koyuncu M, Koluman N, Keskin M, Esenboğa N, Konyalı A, Cemal İ, Gül S, Elmaz O, Koşum N, Dellal G, Bingöl M:** Goat production systems of Turkey: Nomadic to industrial. *Small Rumin Res,* 163, 15-20, 2018. DOI: 10.1016/j.smallrumres.2017.10.001
19. **Devi M, Rahman UH, Weerasinghe WPMCN, Mishra P, Tiwari S, Karakaya K:** Future milk production prospects in India for various animal species using time series models. *Indian J Anim Res,* 56 (9): 1170-1175, 2022. DOI: 10.18805/IJAR.B-4409
20. **Subbanna YB, Kumar S, Puttaraju SKM:** Forecasting buffalo milk production in India: Time series approach. *Buffalo Bull,* 40 (2): 335-343, 2021.
21. **Fernández C, Gómez J, Sánchez-Seiquer P, Gómez-Chova L, Soria-Olivas E, Mocé L, Garcés C:** Prediction of weekly goat milk yield using autoregressive models. *S Afr J Anim Sci,* 34 (Suppl.-1): 169-172, 2004.
22. **Kumar STP, Milind K, Kencharaddi HG, Lahiri B:** Trend and future perspective of milk production in Karnataka. *Indian J Dairy Sci,* 76 (5): 493-498, 2023. DOI: 10.33785/IJDS.2023.v76i05.011
23. **Petcu CD, Gajaila G, Mihai OD, Calin S, Ghimpeteanu OM, Codreanu I, Ciubotaru-Pirvu E:** Research on the influence of lactation stage on goat's milk characteristics. *Sci Papers Ser D Anim Sci,* 65 (1): 536-541, 2022.
24. **Serin H, Siren S:** Examination of meat, milk, and egg production in Turkiye using trend models. *Rev Cient Fac Cienc Vet,* 35 (1): 1-10, 2025. DOI: 10.52973/rcfcv-e35497
25. **Du A, Jia W, Zhang R:** Machine learning methods for unveiling the potential of antioxidant short peptides in goat milk-derived proteins during *in vitro* gastrointestinal digestion. *J Dairy Sci,* 107 (11): 8837-8851, 2024. DOI: 10.3168/jds.2024-24887
26. **Chu C, Wang H, Luo X, Wen P, Nan L, Du C, Fan Y, Gao D, Wang D, Yang Z, Yang G, Liu L, Li Y, Hu B, Abula Z, Zhang S:** Possible alternatives: Identifying and quantifying adulteration in buffalo, goat, and camel milk using mid-infrared spectroscopy combined with modern statistical machine learning methods. *Foods,* 12 (20):3856, 2023. DOI: 10.3390/foods12203856
27. **Lianou DT, Kiouvrekis Y, Michael CK, Vasileiou NGC, Psomadakis I, Politis AP, Katsafadou AI, Katsarou EI, Bourganou MV, Liagka DV, Chatzopoulos DC, Solomakos NM, Fthenakis GC:** The use of explainable machine learning for the prediction of the quality of bulk-tank milk in sheep and goat farms. *Foods,* 13 (24):4015, 2024. DOI: 10.3390/foods13244015
28. **da Silva WST, Silveira RMF, da Silva WE, Leite JHGM, Bermejo LA, Mcmanus C, Krupij AT, Facanha DAE:** Adaptive dynamics of dairy goats in a temperate region. *Int J Biometeorol,* 69 (3): 621-632, 2025. DOI: 10.1007/s00484-024-02843-x
29. **Kaygısız F, Sezgin FH:** Forecasting goat milk production in Turkey using artificial neural networks and box-jenkins models. *Animal Rev,* 4 (3): 45-52, 2017. DOI: 10.18488/journal.ar.2017.43.45.52
30. **Avar MY, Türkoğlu S, Şen C, Işik EE, Ordu E:** Forecasting raw milk supply in Türkiye: a comparative study of machine learning and statistical methods. *5th International Informatics and Software Engineering Conference (IISEC),* 5-6 Feb. Ankara, 2026.
31. **Şahin Demirel AN, Erik T:** Determination of the effect of climate change on small cattle milk yield in Iğdır province via machine learning. *Harran Tarım Gıda Bil Derg,* 28 (3): 374-384, 2024. DOI: 10.29050/harranziraat.1464601
32. **Ordu M, Zengin Y:** A comparative forecasting approach to forecast animal production: A case of Turkey. *Lalahan Hay Araşt,* 60 (1): 24-31, 2020. DOI: 10.46897/lahaed.719095
33. **Duman H:** Application of machine learning algorithms for univariate time series analysis in agricultural forecasting: a case study of cherry production in Turkey. **In,** Köleoğlu N, Çelik Ş, Gülşen İrteş M (Eds): Ekonomi, Sosyal ve Beşeri Bilimlerde İstatistiksel Araştırmalar. Holistence, İstanbul, 2024.

34. **Diebold FX, Mariano RS:** Comparing predictive accuracy. *J Bus Econ Stat,* 13 (3): 253-263, 1995. DOI: 10.1080/07350015.1995.10524599
35. **Harvey DI, Leybourne SJ, Newbold P:** Testing the equality of prediction mean squared errors. *Int J Forecast,* 13 (2): 281-291, 1997. DOI: 10.1016/S0169-2070(96)00719-4
36. **Dickey DA, Fuller WA:** Distribution of the estimators for autoregressive time series with a unit root. *J Am Stat Assoc,* 74 (366): 427-431, 1979. DOI: 10.1080/01621459.1979.10482531
37. **Phillips PCB, Perron P:** Testing for a unit root in time series regression. *Biometrika,* 75 (2): 335-346, 1988. DOI: 10.1093/biomet/75.2.335
38. **Kwiatkowski D, Phillips PCB, Schmidt P, Shin Y:** Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *J Econom,* 54 (1-3): 159-178, 1992. DOI: 10.1016/0304-4076(92)90104-Y
39. **Zivot E, Andrews DWK:** Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis. *J Bus Econ Stat,* 10 (3): 251-270, 1992. DOI: 10.2307/1391541
40. **Lee J, Strazicich MC:** Minimum LM unit root test with two structural breaks. *Rev Econ Stat,* 85 (4): 1082-1089, 2003. DOI: 10.1162/003465303772815961
41. **Lee J, Strazicich MC:** Minimum LM unit root test with one structural break. *Econ Bull,* 33 (4): 2483-2492, 2013.
42. **Yeo IK, Johnson RA:** A new family of power transformations to improve normality or symmetry. *Biometrika,* 87 (4): 954-959, 2000. DOI: 10.1093/biomet/87.4.954
43. **Kuhn M, Johnson K:** Applied Predictive Modeling. Springer, New York, 2013.
44. **Boehmke B, Greenwell B:** Hands-On Machine Learning with R. Chapman and Hall/CRC, New York, 2020.
45. **Nwanganga F, Chapple M:** Practical Machine Learning in R. Wiley, Canada, 2020.
46. **Zhang Z:** Decision tree modeling using R. *Ann Transl Med,* 4 (15):275, 2016. DOI: 10.21037/atm.2016.05.14
47. **Breiman L, Friedman JH, Olshen RA, Stone CJ:** Classification and Regression Trees. Wadsworth, Belmont, CA, 1984.
48. **Friedman JH:** Multivariate adaptive regression splines. *Ann Stat,* 19 (1): 1-67, 1991. DOI: 10.1214/aos/1176347963
49. **Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, Chen K, Mitchell R, Cano I, Zhou T, Li M, Xie J, Lin M, Geng Y, Li Y, Yuan J:** Xgboost: Extreme gradient boosting. *R package,* version 3.4.2.1. https://cran.r-project.org/package=xgboost; *Accessed:* 22.04.2026.
50. **Mishra P, Al Khatib AMG, Yadav S, Ray S, Lama A, Kumari B, Sharma D, Yadav R:** Modeling and forecasting rainfall patterns in India: A time series analysis with XGBoost algorithm. *Environ Earth Sci,* 83 (6):163, 2024. DOI: 10.1007/s12665-024-11481-w
51. **Smola AJ, Schölkopf B:** A tutorial on support vector regression. *Stat Comput,* 14 (3): 199-222, 2004. DOI: 10.1023/B:STCO.0000035301.49549.88
52. **Vapnik V:** The Nature of Statistical Learning Theory. 2nd ed., Springer, New York, 2000.
53. **Thissen U, Van Brakel R, De Weijer AP, Melssen WJ, Buydens LMC:** Using support vector machines for time series prediction. *Chemometr Intell Lab Syst,* 69 (1-2): 35-49, 2003. DOI: 10.1016/S0169-7439(03)00111-4
54. **Box GEP, Jenkins GM:** Time Series Analysis: Forecasting and Control. Revised ed., Holden Day, San Francisco, 1976.
55. **Hyndman RJ, Athanasopoulos G:** Forecasting: Principles and Practice. 3rd ed., OTexts, Melbourne, Australia, 2021. https://otexts.com/fpp3; *Accessed:* 22.05.2026.
56. **Tashman LJ:** Out-of-sample tests of forecasting accuracy: An analysis and review. *Int J Forecast,* 16 (4): 437-450, 2000. DOI: 10.1016/S0169-2070(00)00065-0
57. **Cerqueira V, Torgo L, Mozetič I:** Evaluating time series forecasting models: An empirical study on performance estimation methods. *Mach Learn,* 109 (11): 1997-2028, 2020. DOI: 10.1007/s10994-020-05910-7