

Developing a machine learning prediction model for honey production

Berkant Ismail YILDIZ^{id}, Kemal ESKIOĞLU^{id}, Kemal KARABAG^{id}

Akdeniz University, Faculty of Agriculture, Department of Agricultural Biotechnology, 07058, Konyaaltı, Antalya, Türkiye

Corresponding author: K. Karabag, e-mail: karabag@akdeniz.edu.tr

Author(s) e-mail: berkantyardiz@gmail.com, kemaleskioglu94@gmail.com

ARTICLE INFO

Received: July 6, 2024

Received in revised form: July 17, 2024

Accepted: July 18, 2024

Keywords:

Honey bee

Honey

Production estimate

Artificial intelligence

ABSTRACT

Türkiye, with its rich flora diversity, holds a significant share in global honey production. However, honey bee populations, essential for agricultural ecosystems, face multifaceted threats such as climate change, habitat degradation, diseases, parasites, and exposure to pesticides. Alongside the increasing global food demand driven by population growth, there is a pressing need for a substantial increase in honey production. In this context, advances in machine learning algorithms offer tools to predict future food needs and production levels. The objective of this work is to develop a predictive model using machine learning techniques to predict Türkiye's honey output in the next years. To achieve this goal, a range of machine learning algorithms including K-Nearest Neighbor, Random Forest, Linear Regression, and Gaussian Naive Bayes were employed. Following investigations, Linear Regression emerged as the most effective method for predicting honey production levels ($R^2=0.97$).

1. Introduction

Honey is the main product of beekeeping, which makes a great contribution to the development of rural areas. Honey produced by honey bees is derived largely from flower nectar and is a sweet, natural food transformed by a group of enzymes found in the saliva of worker bees. Honey is aerated to evaporate its water and then stored in hives (Atanasov et al. 2023). Also, honey is one of the most important agricultural products for the Turkish economy. Considered the agricultural beekeeping center of the future due to the richness of its flora (75% of the world's species and varieties) and geographical structure, Türkiye is the world's largest honey producer after China (Coşkun 2019; Atanasov et al. 2023). However, honey bees, which play a critical role in agricultural production and the environment, are threatened by various factors such as climate change, habitat loss, diseases, ecto- and endoparasites and pesticides (Brown et al. 2016; Potts et al. 2016; Pătruică et al. 2021). Biotic and abiotic factors that affect the health status of honey bees also affect honey production (Olate-Olave et al. 2021). Previous studies have shown that climate parameters such as temperature, humidity and rainfall affect honey production and honey quality (Oroian et al. 2017; Clarke and Robert 2018; Fatima et al. 2022; Şengül et al. 2023). In addition, honey bee losses caused by pesticide use and the decrease in foraging activity, resulting from diminishing the amount of agricultural and forest areas, lead to a decrease in honey production (Ferreira et al. 2015; Alqarni et al. 2021; Abay et al. 2023).

The exponential and unregulated growth of the global population will lead to a corresponding surge in the demand for food. For instance, it is stated that world food production must increase by approximately 60% to feed the world population in 2050 (Van Dijk et al. 2021). When we look at Türkiye specifically, it is expected that the country's population will reach 100.4 million by the year 2050 (TÜİK 2023). Considering the

current population of 85 million, at least a 20% increase in honey production is required to maintain the supply-demand balance. Given the challenges honey bees are now facing, it is crucial to enhance honey output and productivity in order to fulfill the future need for food.

Obtaining data on agricultural production before harvest provides significant advantages in both obtaining information about the production process and achieving sustainable development goals. In production estimates, reliable statistical procedures such as multivariate statistical methods are generally applied (Niazian and Niedbala 2020). Nowadays, with the rapid development of technology, approaches such as machine learning algorithms (MLA) based on statistical methods are widely used for production estimates and similar business processes (Ahmed and Hussain 2022). Machine learning (ML) is an important sub-branch of artificial intelligence, and its main purpose is to be able to handle complexities in large data sets and make such estimations because of their learning abilities (Kononenko 2001). While there is limited knowledge regarding the interaction of factors influencing honey production and their impact on honey production, one of the most significant approaches in this regard is the evaluation of honey production estimation within a ML model.

Looking at the literature, although a lot of work has been done on various processes in different species using machine learning, there are few studies on honey bees. Nevertheless, studies have shown that machine learning techniques are quite suitable and useful for analyzing beekeeping data. Prešern and Smodiš Škerl (2019) used the Gradient Boosting Machine algorithm of the machine learning software H2O to estimate the parameters affecting queen body mass. They developed three different models using different parameter combinations and in

their results, they determined that "ovary mass" and "breeder" parameters are the most important factors in model estimations. [Campbell et al. \(2020\)](#) used machine learning to evaluate the capacity of both weather data and satellite-derived vegetation data to develop a predictive model for Marri honey harvest in South Western Australia. Regression Trees were able to predict Marri honey harvested per hive to a Mean Error (MAE) of 10.3 kg. [Calovi et al. \(2021\)](#) aimed to evaluate the importance of weather, topography, land use and management factors on winter mortality in honey bee colonies using the Random Forest algorithm and to estimate survival given the existing factors. Random Forest estimated overwintering survival with 73.3% accuracy for colonies and 65.7% accuracy for beehives with managed Varroa mite populations. Additionally, growing degree days and precipitation in the warmest quarter of the previous year were the most important determinants at both levels. [Veiner et al. \(2022\)](#) tested three supervised learning algorithms (Random Forests, Lasso and Elastic net Regularized Generalized Linear Model, and Support Vector Machine) for their performance in characterizing transcriptomic patterns and identifying genes associated with honey bee waggle dance. By matching the analysis results with differential gene expression outputs, they identified two candidate genes for the neural regulation of waggle dance. [Braga et al. \(2023\)](#) developed a machine learning model to estimate temperature drops in honey bee colonies, the Long Short-Term Memory algorithm, which was applied to five real data sets with input factors of internal temperature, internal humidity, mean fanning, mean noise, mass and external temperature. Their results showed that they could predict the temperature 24 hours in advance with a Root Mean Square Error (RMSE) of 0.5%.

Previous studies have explored the impact of some environmental factors and climatic conditions on honey production ([Ferreira et al. 2015](#); [Oroian et al. 2017](#); [Clarke and Robert 2018](#); [Alqarni et al. 2021](#); [Fatima et al. 2022](#); [Abay et al. 2023](#); [Şengül et al. 2023](#)). Drawing on the information obtained from these studies, this study attempted to develop a predictive model using machine learning techniques to predict Türkiye's honey output over the next years.

2. Material and Methods

2.1. Material

This study incorporates eight distinct attributes: honey production volume, number of enterprises, number of colonies, pesticide application, agricultural and forest land coverage, as well as temperature and rainfall. The selection of these attributes was informed by their relevance to honey production dynamics. Given the systematic recording of agricultural data in Türkiye since the early 2000s, this study utilized annual data collected between 2000 and 2022 as the TrbalDataSet training set. Agricultural data was sourced from the Republic of Türkiye Ministry of Agriculture and Forestry website, while meteorological data was obtained from the Turkish State Meteorological Service website. There is no null value in the data set. In the study, the min-max normalization method was applied to normalize the data set values between 0 and 1.

2.2. Methods

The study employed the Python programming language, along with popular libraries such as Pandas, Numpy, Matplotlib, and Scipy, for coding machine learning algorithms. A 75% portion of the dataset was allocated for the training phase of each

algorithm, while the remaining 25% served as the test set to assess the accuracy of predictions derived from this training.

The creation of estimation models involved the selection of commonly used machine learning methods, outlined below:

- K-Nearest Neighbors (KNN): KNN is a non-parametric supervised learning algorithm employed in classification and regression domains. Its popularity has surged across diverse fields due to its simplicity of application and reliance on a straightforward mathematical foundation. The algorithm identifies the nearest neighbors to a given point and utilizes these points for estimations. At its core, KNN operates on the concept that the outcome of an event mirrors that of its closest neighboring events. The parameter "K" signifies the number of closest points considered in the estimation process ([Hai et al. 2023](#)).

- Random Forest (RF): This algorithm operates as an ensemble learning method founded on decision trees, a type of supervised learning technique commonly utilized in machine learning to construct estimation models. Specifically, it forms a random forest, which is an assembly of decision trees trained through the bagging method. The rationale behind employing the bagging method is to enhance overall results by combining multiple learning models ([Breiman 2001](#)). The Random Forest approach proves highly effective when dealing with datasets featuring a multitude of predictors, especially in cases where variable relationships are nonlinear or intricate. This is because it offers flexibility and is not constrained by specific distributions ([Shoemaker et al. 2018](#)).

- Linear Regression (LR): Linear Regression stands as a machine learning technique employed to quantify the association between two variables. The objective of this method is to model the linear correlation between the independent variable (x) and the dependent variable (y). Its primary aim is to predict the dependent variable based on the values of the independent variables ([Maulud and Abdulazeez 2020](#)).

- Gaussian Naive Bayes (GNB): Developed on the principles of Bayes theorem, Naive Bayes stands out as one of the most straightforward, comprehensible, and practical machine learning algorithms employed in classification tasks. The term "naive" is attributed to the algorithm due to its assumption of independence among features during the classification process ([John and Langley 2013](#)).

The test set was contrasted with the values generated by the algorithms' estimations, and the evaluation encompassed the coefficient of determination (R^2) along with error metrics, including Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and Mean Absolute Error (MAE) ([Rahman et al. 2021](#)). Previous studies ([Gültepe 2019](#); [Rahman et al. 2021](#)) indicate that algorithms exhibiting R^2 values nearing 1 and error values approaching 0 are considered the most effective.

3. Results and Discussion

In tests conducted with machine learning, metrics such as R^2 , MSE, RMSE and MAE were taken into account in evaluating the performance of the algorithms. Accordingly, in [Table 1](#), the lowest RMSE (2966.83) value among the algorithms for honey production estimation was calculated for LR. Similarly, in terms of MAE (2455.34), MAPE (0.03), MedAE (2953.96) values, the lowest was calculated for the LR algorithm. Another important metric that measures the performance of the ML model is R^2 .

Table 1. Comparison of the results of machine learning algorithms used on TrbalDataSet

Algorithms	R ²	RMSE	MAE	MAPE	MedAE
K-Nearest Neighbors (KNN)	0.60	11635.88	9265.17	0.10	6998.50
Random Forest (RF)	0.58	11864.28	9678.50	0.10	8238.50
Linear Regression (LR)	0.97	2966.83	2455.34	0.03	2953.96
Gauss Naive Bayes (GNB)	0.60	11635.88	9265.17	0.10	6998.50

* R², coefficient of determination; RMSE, Root Mean Square Error; MAE, Mean Absolute Error; MAPE, Mean Absolute Percentage Error; MedAE, median absolute error.

Looking at the table 1, R² values for KNN, RF, LR and GNB are calculated as 0.60, 0.58, 0.97, 0.60 respectively. The highest R² value was obtained for LR, indicating higher accuracy.

After data learning, the regression graphs obtained for trains and tests are shown in Figure 1. The obtained graphs demonstrate the concordance between the algorithm and real data. When examining the graphs, it is evident that the Linear Regression algorithm is the most compatible with the real data.

According to the scores obtained, the Linear Regression method showed the highest performance and provided convincing evidence that annual honey production estimation is possible. When we estimate honey production in 2050 with the data between 2000 and 2022 using this algorithm, it is expected that 218,271.34 tons of production will occur. This amount is more than 60% of the 2022 honey production amount and meets the previously mentioned food production increase required by 2050 (Van Dijk et al. 2021). Honey production is affected by environmental factors (Ferreira et al. 2015; Oroian et al. 2017; Clarke and Robert 2018; Alqarni et al. 2021; Fatima et al. 2022; Abay et al. 2023; Şengül et al. 2023). Suitable areas are required to carry out beekeeping activities, and forest areas are ideal regions for beekeeping in terms of climate conditions and vegetation (Güngör and Ayhan 2016). In particular, the honey forest action plan, which was implemented by the Türkiye Ministry of Agriculture and Forestry in 2013 to increase honey production and is still ongoing (Karaağaç and Bulut 2023), is thought to have had a significant impact on the increase in the honey production amount, number of enterprises, number of colonies, and amount of forest area over the last 10 years. Considering these effects, similar projects should be increased in the coming years and Türkiye's beekeeping potential should be utilized more. Thus, by increasing honey production, positive results will be achieved in terms of food security.

No prior studies have been identified in Türkiye where the quantity of honey production was estimated using machine learning techniques. While various traditional statistical methods have been employed to estimate honey production, these approaches provide shorter-term and comparatively less reliable predictions. For instance, Burucu and Gülse Bal (2017) conducted a 7-year (2017-2023) estimation using the ARIMA model with data sourced from TURKSTAT, projecting a continuous increase in honey production in Türkiye, reaching 121,216 tons in 2023. In a separate study, the ARIMA model was utilized to predict honey supply and demand in Türkiye for the period 2016-2023. The study estimated honey supply per capita to be 1.43 kg in 2017 and 1.54 kg in 2023 (Naseri et al. 2016). Çukur and Çukur (2021) aimed to estimate the quantity of honey production in Türkiye using the Box Jenkins ARIMA model, incorporating honey production data from 1990-2019. The results suggested an estimated honey production of 123420 tons in the year 2025. However, these models tend to produce

inaccurate estimations when dealing with multidimensional input data. Recognizing this limitation, machine learning is now widely adopted, offering highly accurate predictions for complex activities, such as agricultural production, by considering numerous ecological variables (Rahman et al. 2021).

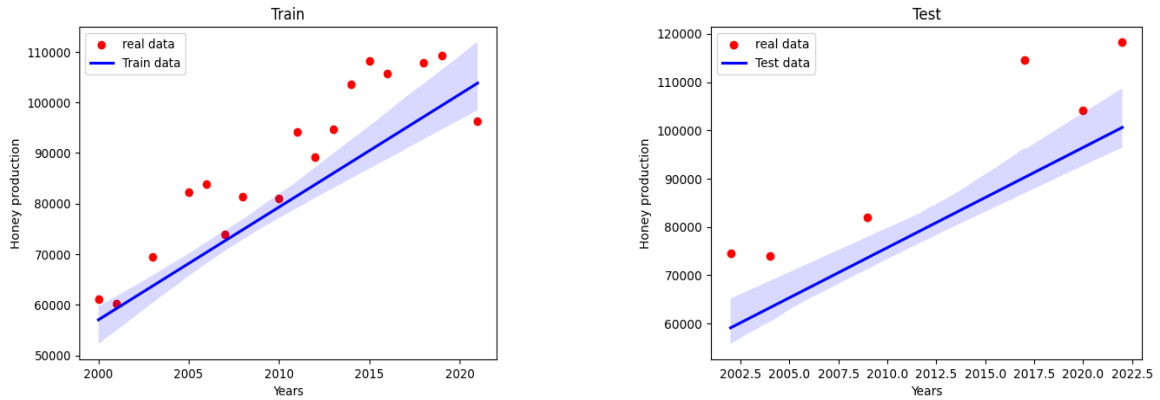
Nevertheless, the pervasive issue of climate change poses a significant challenge to the accuracy of future honey production estimates. Climate change has far-reaching implications for ecological systems and honey bee populations, being closely linked to colony collapse disorder (Pătruică et al. 2021; Şengül et al. 2023). The adverse effects on pollination activities and overall productivity, coupled with the potential increase in infectious diseases and parasites, such as Varroa destructor, highlight the vulnerability of honey bee colonies (Klein et al. 2007; Switaneck et al. 2015). Unfortunately, these challenges are expected to intensify in the years to come (Varol and Yücel 2019). To mitigate these risks, various adaptation strategies have been proposed, including reforestation, hive sterilization improvement, queen bee replacement, artificial feeding, breeding, migratory beekeeping, honey crop cultivation, changes in apiary management, adoption of good beekeeping practices, seeking technical assistance, and maintaining comprehensive records. However, obstacles such as insufficient funding, limited availability of suitable beekeeping land, and bureaucratic challenges in migratory beekeeping impede the widespread adoption of these strategies by beekeepers (Şengül et al. 2023).

4. Conclusions

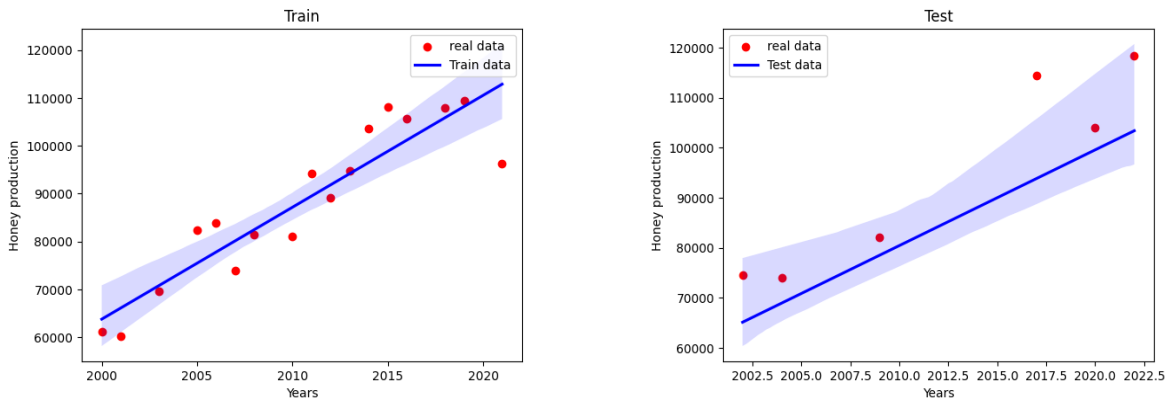
This study has utilized machine learning to create a prediction model for Türkiye's honey production, offering a fresh perspective to help understand the intricate dynamics of honey production. Despite the reliability of the metrics obtained from our analysis, it is advisable to consider additional production dynamics to offer a more comprehensive outlook for the upcoming years. Nevertheless, shortcomings exist in beekeeping production data within Türkiye, particularly in health records.. To improve production forecasts and realistically portray food supply and demand in the future, a meticulous record-keeping approach is essential. Moreover, it is crucial to embrace climate change adaptation strategies and to advocate for these strategies to be part of agricultural policy in order to sustain economic activities in beekeeping, especially honey production, in Türkiye.

In summary, while machine learning, specifically the Linear Regression method, proves invaluable in refining the accuracy of honey production estimates, the urgent challenges presented by climate change highlight the necessity for ongoing research, innovation, and collaborative endeavours. These efforts are critical for ensuring the sustainability of beekeeping practices and addressing food security concerns in the face of evolving environmental dynamics.

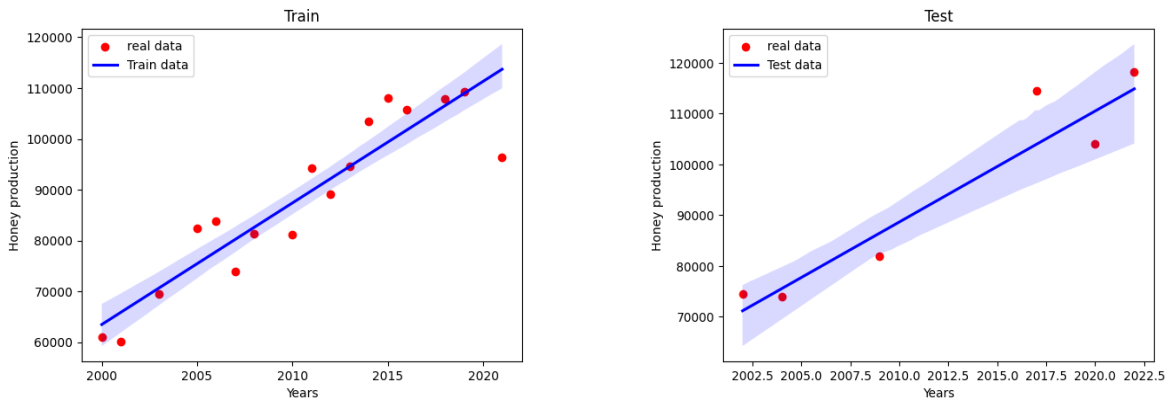
a)



b)



c)



d)

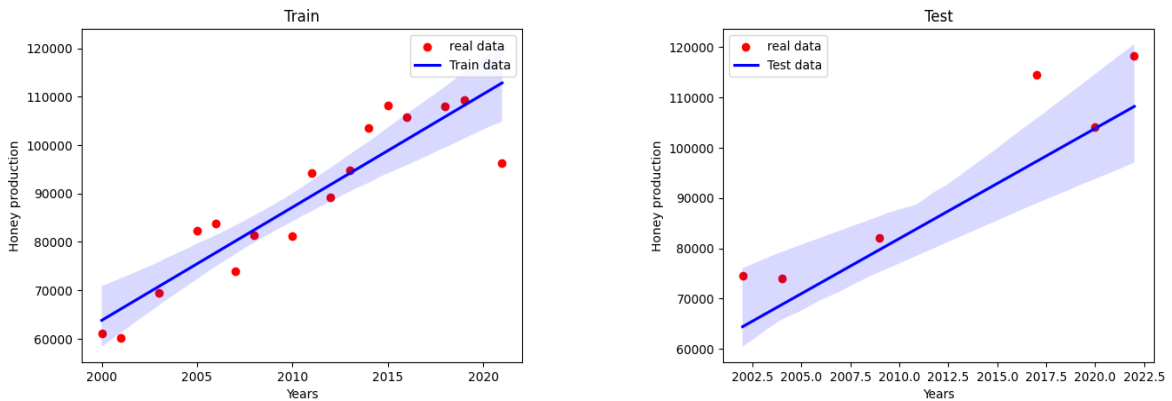


Figure 1. The train and test regression graphs obtained from the algorithms. The red dots represent real data, while the blue curves represent estimated data. a) KNN train and test graph b) RF train and test graph c) LR train and test graph d) GNB train and test graph.

References

- Abay Z, Bezabeh A, Gela, A, Tassew A (2023) Evaluating the Impact of Commonly Used Pesticides on Honeybees (*Apis mellifera*) in North Gonder of Amhara Region, Ethiopia. *Journal of Toxicology* 2023: 2634158. doi: 10.1155/2023/2634158.
- Ahmed MU, Hussain I (2022) Prediction of wheat production using machine learning algorithms in northern areas of Pakistan. *Telecommunications Policy* 46: 102370. <https://doi.org/10.1016/j.telpol.2022.102370>.
- Alqarni AS, Iqbal J, Raweh HS, Hassan A, Oways AA (2021) Beekeeping in the Desert: Foraging activities of honey bee during major honeyflow in a hot-arid ecosystem. *Applied Science* 11: 9756. doi: 10.3390/app11209756.
- Atanasov AZ, Georgiev SG, Vulkov LG (2023) Parameter estimation analysis in a model of honey production. *Axioms* 12: 214. doi: 10.3390/axioms12020214.
- Braga AR, Freitas BM, Gomes DG, Bezerra AD, Cazier JA (2021) Forecasting sudden drops of temperature in pre-overwintering honeybee colonies. *Biosystems Engineering* 209: 315-321. doi: 10.1016/j.biosystemseng.2021.07.009.
- Breiman L (2001) Random forests. *Machine Learning* 45: 5-32. doi: 10.1023/A:1010933404324.
- Brown MJ, Dicks LV, Paxton RJ, Baldock KC, Barron AB, Chauzat MP, Freitas BM, Goulson D, Jepsen S, Kremen C, Li J, Neumann P, Pattermore DE, Potts SG, Schweiger O, Seymour CL, Stout JC (2016) A horizon scan of future threats and opportunities for pollinators and pollination. *PeerJ* 4: e2249. doi: 10.7717/peerj.2249.
- Burucu V, Gülse Bal HS (2017) Türkiye’de arıcılığın mevcut durumu ve bal üretim öngörüsü. *Tarım Ekonomisi Araştırmaları Dergisi* 3: 28-37.
- Calovi M, Grozinger CM, Miller DA, Goslee SC (2021) Summer weather conditions influence winter survival of honey bees (*Apis mellifera*) in the northeastern United States. *Scientific Reports* 11: 1553. doi: 10.1038/s41598-021-81051-8.
- Campbell T, Dixon KW, Dods K, Fearn P, Handcock R (2020) Machine learning regression model for predicting honey harvests. *Agriculture* 10: 118. doi: 10.3390/agriculture10040118.
- Clarke D, Robert D (2018) Predictive modelling of honey bee foraging activity using local weather conditions. *Apidologie* 49: 386-396. doi: 10.1007/s13592-018-0565-3.
- Coşkun A (2019) Türkiye’de bal sektörünün mevcut durum değerlendirilmesi ve tüketici eğilimleri. Yüksek Lisans Tezi, Namık Kemal University, Tekirdağ.
- Çukur T, Çukur F (2021) ARIMA modeli ile Türkiye bal üretim öngörüsü. *Tarım Ekonomisi Araştırmaları Dergisi* 7(1): 31-39.
- Fatima N, Mujtaba G, Akhuzada A, Ali K, Shaikh ZH, Rabia B, Shahani JA (2022) Prediction of Pakistani honey authenticity through machine learning. *IEEE Access* 10: 87508-87521.
- Ferreira PA, Boscolo D, Carvalheiro LG, Biesmeijer JC, Rocha PL, Viana BF (2015) Responses of bees to habitat loss in fragmented landscapes of Brazilian Atlantic Rainforest. *Landscape Ecology* 30: 2067-2078.
- Gültepe Y (2019) Makine öğrenmesi algoritmaları ile hava kirliliği tahmini üzerine karşılaştırmalı bir değerlendirme. *Avrupa Bilim ve Teknoloji Dergisi* 16: 8-15. doi: 10.31590/ejosat.530347.
- Güngör E, Ayhan A (2016) Bartın yöresi orman kaynaklarının bal üretim potansiyeli ve ekonomik değeri. *Turkish Journal of Forestry* 17: 108-116. doi: 10.18182/tjf.89126.
- Hai A, Bharath G, Patah MFA, Daud WMAW, Rambabu K, Show P, Banat F (2023) Machine learning models for the prediction of total yield and specific surface area of biochar derived from agricultural biomass by pyrolysis. *Environmental Technology & Innovation* 30: 103071. doi: 10.1016/j.eti.2023.103071.
- John GH, Langley P (2013) Estimating continuous distributions in Bayesian classifiers. *arXiv preprint arXiv:1302.4964*. doi: 10.48550/arXiv.1302.4964.
- Karaağaç S, Bulut İ (2023) Antalya’daki bal ormanlarının dağılımı ve sürdürülebilir yönetimi (2010-2022). *Akdeniz Üniversitesi Sosyal Bilimler Enstitüsü Dergisi* 13: 56-73.
- Klein AM, Vaissiere BE, Cane JH, Steffan-Dewenter I, Cunningham SA, Kremen C, Tscharntke T (2007) Importance of pollinators in changing landscapes for world crops. *Proceedings of the Royal Society B: Biological Sciences* 274: 303-313. doi: 10.1098/rspb.2006.3721.
- Kononenko I (2001) Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine* 23: 89-109. doi: 10.1016/s0933-3657(01)00077-x.
- Maulud D, Abdulazeez AM (2020) A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends* 1: 140-147. doi: 10.38094/jastt1457.
- Naseri Z, Saner G, Adanacıoğlu H (2016) The Future Trends of Honey Supply and Demand in Turkey (OR-70). In: 5th International Mugla Beekeeping and Pine Honey Congress, Healthy Bees-Healthy Life. Mugla, Türkiye, pp. 200-201.
- Niazian M, Niedbala G (2020) Machine learning for plant breeding and biotechnology. *Agriculture* 10: 436. doi: 10.3390/agriculture10100436.
- Olate-Olave VR, Verde M, Vallejos L, Perez Raymonda L, Cortese MC, Doorn M (2021) Bee health and productivity in *Apis mellifera*, a consequence of multiple factors. *Veterinary Science* 8: 76. doi: 10.3390/vetsci8050076.
- Oroian M, Ropciuc S, Paduret S, Sanduleac ET (2017) Authentication of Romanian honeys based on physicochemical properties, texture and chemometric. *Journal of Food Science and Technology* 54: 4240-4250. doi: 10.1007/s13197-017-2893-0.
- Pătruică S, Pet I, Simiz E (2021) Beekeeping in the context of climate change. *Scientific Papers. Series D. Animal Science* 64: 2393-2260.
- Potts SG, Imperatriz-Fonseca V, Ngo HT, Aizen MA, Biesmeijer JC, Breeze TD, Dicks LV, Garibaldi LA, Hill R, Settele J, Vanbergen AJ (2016) Safeguarding pollinators and their values to human well-being. *Nature* 540: 220-229. doi: 10.1038/nature20588.
- Prešern J, Smodiš Škerl MI (2019) Parameters influencing queen body mass and their importance as determined by machine learning in honey bees (*Apis mellifera carnica*). *Apidologie* 50: 745-757. doi: 10.1007/s13592-019-00683-y.
- Rahman LF, Marufuzzaman M, Alam L, Bari MA, Sumaila UR, Sidek LM (2021) Developing an ensemble machine learning prediction model for marine fish and aquaculture production. *Sustainability* 13: 9124. doi: 10.3390/su13169124.
- Şengül Z, Yücel B, Saner G, Takma Ç (2023) Investigating the Impact of Climate Parameters on Honey Yield under Migratory Beekeeping Conditions through Decision Tree Analysis: The Case of İzmir Province. *ANADOLU Ege Tarımsal Araştırma Enstitüsü Dergisi* 33: 268-280. doi: 10.18615/anadolu.1394787.
- Shoemaker KT, Heffelfinger LJ, Jackson NJ, Blum ME, Wasley T, Stewart KM (2018) A machine-learning approach for extending classical wildlife resource selection analyses. *Ecology and Evolution* 8: 3556-3569. doi: 10.1002/ece3.3936.
- Switaneck M, Brodschneider R, Crailsheim K, Truhetz H (2015) Impacts of austrian climate variability on honey bee mortality In: EGU General Assembly Conference. Vienna, 17, 9575.
- TÜİK (2023) Nüfus projeksiyonları. <https://data.tuik.gov.tr/Bulten/Index?p=Nufus-Projeksiyonlari-2018-2080-30567>. Accessed 22 March, 2024.
- Van Dijk M, Morley T, Rau ML, Saghai Y (2021) A meta-analysis of projected global food demand and population at risk of hunger for the period 2010–2050. *Natur Food* 2: 494-501. doi: 10.1038/s43016-021-00322-9.

- Varol E, Yücel B (2019) The effects of environmental problems on honey bees in view of sustainable life. *Mellifera* 19: 23-32.
- Veiner M, Morimoto J, Leadbeater E, Manfredini F (2022) Machine learning models identify gene predictors of waggle dance behaviour in honeybees. *Molecular Ecology Resources* 22: 2248-2261. doi: 10.1111/1755-0998.13611.