Classification of Brand Perception Using Random Forest: Brand Preference, Brand Loyalty, and Brand Trust

Muhammed İkbal Yılmaz[1*+], Özge Yıldız[2] and Şahika Gökmen[3,4]

[1, 2] Undergraduate Student, Department of Econometrics, Faculty of Economics and Administrative Sciences, Hacı Bayram Veli University, 06500, Ankara, Turkey.

[3] Associate Professor, Department of Econometrics, Faculty of Economics and Administrative Sciences, Hacı Bayram Veli University, 06500, Ankara, Turkey.

[4] Researcher, Statistics Department, Uppsala University, Uppsala, Sweden.

[* +] *Corresponding author and Speaker e-mail: myucanlar@gmail.com*

Presentation/Paper Type: Oral

Abstract

Brands enable strategy development, competitive advantage, and audience connection in modern society. This study addresses gaps in Brand Preference (BP), Brand Loyalty (BL), and Brand Trust (BT) among university students using machine learning, particularly Random Forest (RF) Classification. Data from Ankara Hacı Bayram Veli University (Spring 2023/2024) is analyzed with decision tree classifiers. A meta-predictor reduces overfitting, while the 'Brand Perception Scale', with 22 independent variables (x_i), assesses BP, BL, and BT.

Keywords*: Machine Learning Classification, Brand Trust, Brand Loyalty, Brand Preference, Random Forest, Decision Trees*

INTRODUCTION

In AI, McCulloch et al. (1943) introduced the artificial neuron model, laying the foundation for Deep Learning (DL). Turing's (1950) query, "Can Machines Think?" hinted at machine emulation of human responses. On August 31, 1955, McCarthy and colleagues coined "AI" while defining machine learning capabilities. Samuel (1959) later defined ML as a computer's ability to improve from experience. Initially, brands served as simple identifiers (Davis, 1964) but have since evolved into complex assets essential for growth, rooted in 'brand image' factors like quality and reputation (Aaker, 1996; Keller, 2001). BT is linked to belief and behavior (Moorman et al., 1992), while early identification of BL factors strengthens brand ties (Schee, 2010). Key drivers of BL include quality (Zehir et al., 2011), and brands have become complex, influential assets (Bastos and Levy, 2012). BP reflects individual preferences shaped by

multiple factors (Ebrahim, 2013; Yang et al., 2015). While research on BL clustering is limited, BL aligns with repurchasing loyalty (Gumus, 2016; Li and You, 2021). BT has become critical for customer loyalty and long-term relationships (Lee et al., 2014; Alhaddad, 2015), with technological innovation as a strategic asset (Morgan-Thomas and Veloutsou, 2013). Despite extensive ML-brand research (Chaudhary et al., 2016; Saylı et al., 2016; Pamuksuz and Yun, 2021; Dong, 2023), studies integrating BP, BL, and BT are rare. To address this, we applied a RF multi-class classification model to clarify relationships among these constructs.

DECISION TREES WITH MACHINE LEARNING

Decision Trees

This study employed proportional quota sampling to select a representative sample of 407 undergraduate students from the Econometrics and Economics departments at Ankara Hacı Bayram Veli University, each exposed to brands. The dataset was split into a 75% training and 25% test set, revealing significant class imbalance (Figure 1), which complicates classification and necessitates adjustments to enhance predictive accuracy. Chawla et al. (2002) demonstrated the efficacy of SMOTE (Synthetic Minority Over-sampling Technique) for improving minority class prediction in imbalanced datasets. Pears et al. (2014) observed that while synthetic data generation can balance classes, it may risk overfitting and potential information loss. SMOTE mitigates these risks by interpolating within the minority class data, ensuring synthetic samples align with the original data without introducing external information. Here, x_i represents a minority class sample, with J selected via the k-nearest neighbors method (Tesfahun and Bhaskari, 2013). In this approach, λ is a random variable within $[0, 1]$, and the difference vector is $\Delta = J - x_i$. A new synthetic data point is generated as $y_{\text{new}} = x_i + \lambda \cdot \Delta$, creating a sample between x_i and J. If $\lambda = 1$, y_{new} matches J; if $\lambda = 0$, y_{new} remains x_i. Quinlan (1986) made key contributions to decision tree algorithms, essential for RF classification, which begins with tree construction. In RF, data and variables are randomly sampled to construct M decision trees, each based on unique training dataset subsets: $D_n^{(j)} = \{(x_1, y_1^{(j)}), (x_2, y_2^{(j)}), \ldots, (x_n, y_n^{(j)})\}$, where $y_i^{(j)} = h(x_i; \theta_j, D_n^{(j)})$. Each tree $h(x_i; \theta_j, D_n^{(j)})$ predicts the class of x_i using tree-specific split criteria. The j-th tree's prediction is $y_i^{(j)}$. Final classification of x_i is determined by majority vote across M trees, with the highest-voted class selected as: $\hat{y}(x_i) = \arg\max_{c \in C} \left\{ \sum_{j=1}^{M} 1 \left[h(x_i; \theta_j, D_n^{(j)}) = c \right] \right\}$ where C denotes all possible classes (As given in Table 1), and $h(x_i; \theta_j, D_n^{(j)})$ represents tree j's prediction of x_i based on parameters θ_j. This ensemble approach combines multiple trees for a robust model (Kursa, 2014).

Machine Learning

Data analysis was conducted in Python $3.12.5$ (Windows 64-bit) on Jupyter Notebook, using a system with an 11th Gen Intel Core i5-$11260H$, 24 GB RAM, and NVIDIA GeForce RTX 2050ti GPU (Micro-Star International Co.). Key libraries included `pandas` for data handling, `numpy` for computations, `imb-learn.over_sampling.SMOTE` for class imbalance, and `sklearn.ensemble.RandomForestClassifier` for classification modeling, with `sklearn.metrics.classification_report` and `sklearn.metrics.-confusion_matrix` for evaluation.

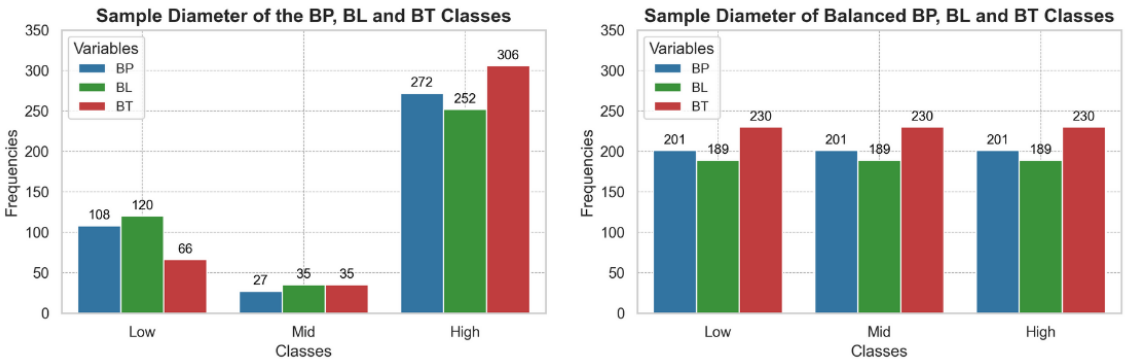

Figure 1. Original Dataset and Balanced Dataset

BP, BL, and BT models were based on averages from 7, 4, and 3 questions, respectively. Internal consistency (Cronbach's Alpha, α) values were 0.7417 for BP, 0.7821 for BL, and 0.7696 for BT, with an overall reliability of $\alpha = 0.8399$.

Table 1. Variables and Classes

Variable Type	Variables	Classes
Dependent (y_i)	BP	Low, Mid, High
	BL	Low, Mid, High
	BT	Low, Mid, High
Independent (x_i)	Age, Gender, Department, Year, BandR, Place, Income, Social media, Ishopping, Brand Monitoring, Price, Quality, Time of use, Production Date, Seasonality, Material, Discount, Fashion, Design, Psychological impulses, Country, Recognizability	

Table 1 details y_i and x_i in the dataset. Three classification models categorized observations into Low, Mid, and High classes using the RF algorithm with Breiman's bootstrap resampling on training set D_n ($D_n^{(j)} \sim \text{Bootstrap}(D_n)$) and a voting mechanism (Breiman, 1996). Each model included 22 independent variables. RF, introduced by Breiman (1984; 2001), is a prominent ML algorithm advancing decision tree methods (Quinlan, 1986). Table 2 presents classification performance pre- and post-SMOTE, showing BP, BL, and BT model differences. BP achieved high Precision (0.83) and Recall (0.78), while BL and BT had Recalls of 0.74 and 0.70. BP and BL excelled in the High class (Recalls 0.88 and 0.91), and both BP and BT showed balanced Precision and Recall in the Low class (0.82/0.89 for BP; 0.85/0.81 for BT).

Table 2. Performance Metrics for Imbalanced and Balanced Data

Variables	Classes	Precision	Recall	F1-Score	Variables	Classes	Precision	Recall	F1-Score
	Low	0.73	0.42	0.54		Low	0.82	0.89	0.85
$\hat{y}_{BP}$	Mid	1.00	0.00	0.00	$\hat{y}_{BP}$	Mid	0.83	0.78	0.80
	High	0.77	0.94	0.85		High	0.91	0.88	0.89
	Accuracy:	0.76				**Accuracy:**	0.85		
	Low	0.80	0.48	0.60		Low	0.80	0.83	0.81
$\hat{y}_{BL}$	Mid	1.00	0.00	0.00	$\hat{y}_{BL}$	Mid	0.86	0.74	0.79
	High	0.73	0.95	0.83		High	0.82	0.91	0.86
	Accuracy:	0.75				**Accuracy:**	0.82		
	Low	1.00	0.31	0.48		Low	0.85	0.81	0.83
$\hat{y}_{BT}$	Mid	1.00	0.00	0.00	$\hat{y}_{BT}$	Mid	0.81	0.70	0.75
	High	0.78	1.00	0.88		High	0.75	0.87	0.80
	Accuracy:	0.79				**Accuracy:**	0.80		

Figure 2's confusion matrices display model predictions on the balanced dataset. BP improved with some Mid and High reassignments to Low. BL accuracy increased with refined Low and

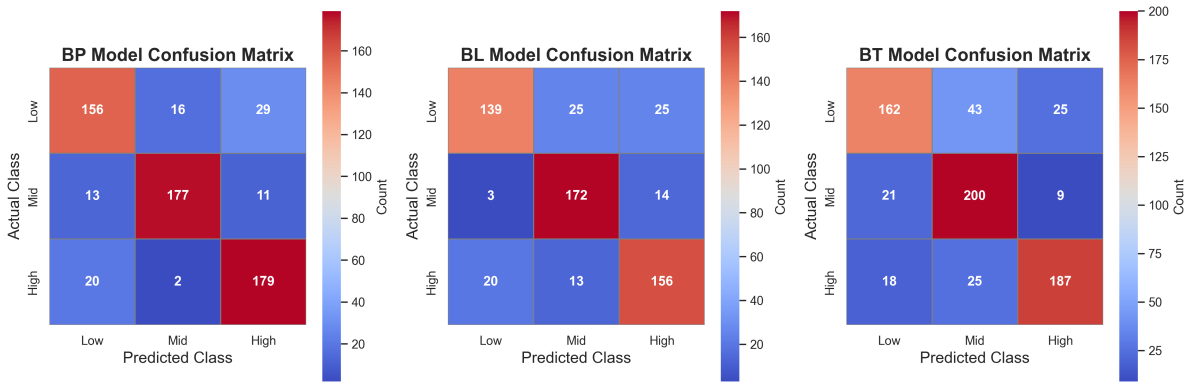

Figure 2. Confusion Matrices

High predictions, while BT showed strong Mid class performance. Adding High and Mid confidence students to Low BT supported model balance. Figures 3, 4, and 5 show decision trees where entropy $H(X)$ at each node measures system uncertainty (Shannon, 1948; Tian et al., 2019). Social Media (≤ 3.5) is a primary BP driver, especially affecting Mid and High classes along with Fashion. Low social media use but high Quality (≤ 4.5) links to stronger BP, while low Brand Monitoring (≤ 1.5) and Seasonality (≤ 2.5) link to Mid BP. Lower Fashion (≤ 3.5) and Design (≤ 4.5) scores associate with Low BP. Findings show Fashion and Social Media significantly influence BP: students engaged with fashion, social media, or higher income tend toward High BP, while less engaged students align with Mid or Low BP. Financial awareness also impacts BP, especially for low social media users (Figure 3). The BL decision tree in Fig-

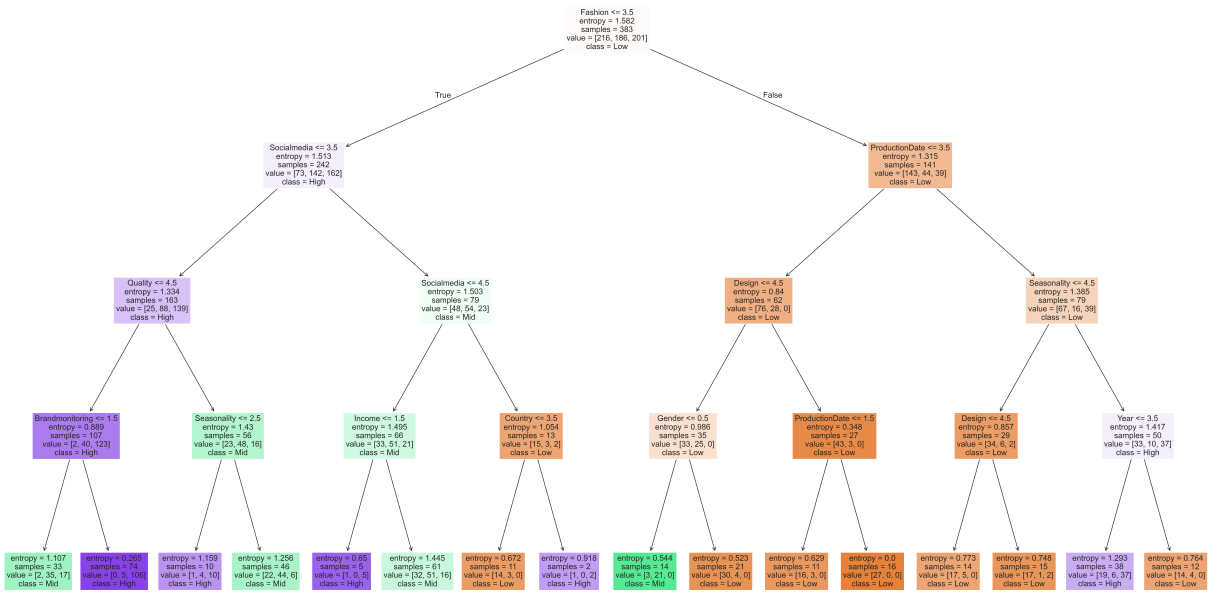

Figure 3. BP Model Decision Tree

ure 4 identifies design perception as the main factor influencing BL. Students with lower design appreciation tend toward Low BL, while higher ratings align with Mid or High BL. Online shopping interactions boost loyalty, though this effect declines for older products. Production recency positively impacts BL, with discount perceptions further boosting loyalty, especially for students active on social media. Recognizability and income also influence BL, with higher incomes linked to greater loyalty. Key BL factors include design, shopping behavior, and production recency, while low design interest and online engagement relate to Low BL.

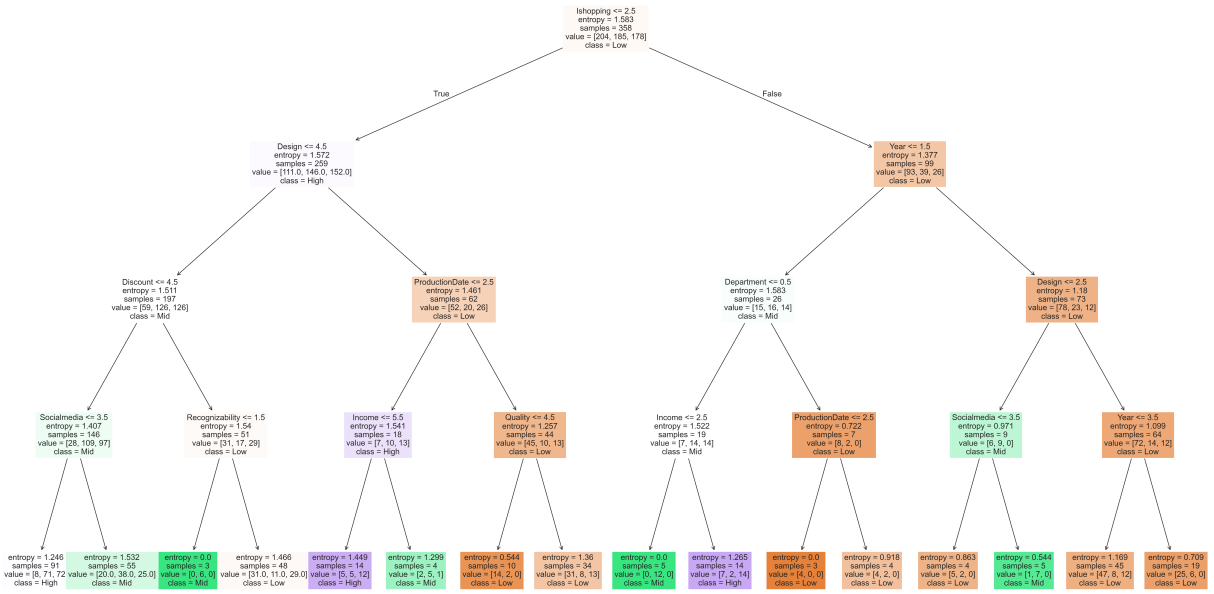

Figure 4. BL Model Decision Tree

The BT decision tree in Figure 5 shows online shopping as the top trust factor, with lower engagement linked to Low BT and higher engagement showing varied trust levels. Design perception also impacts BT: low appreciation aligns with Low BT, high appreciation with Mid or High BT. Recognizability strengthens trust, especially for high-income students, while Low

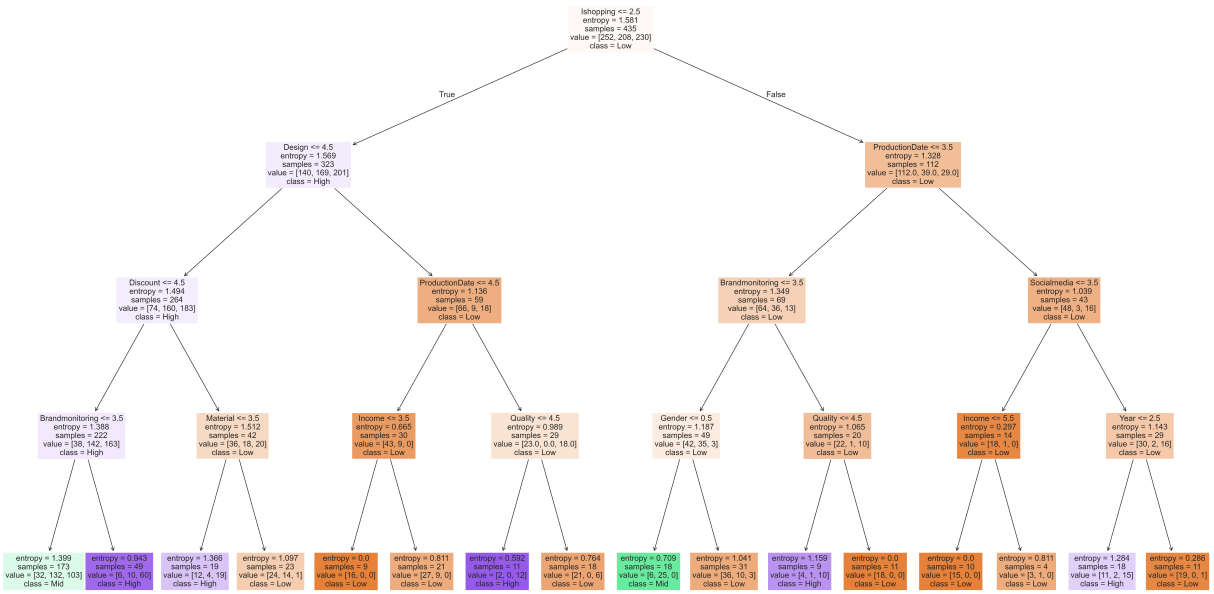

Figure 5. BT Model Decision Tree

Income students tend to remain in Low BT. Ignoring production dates and design lowers trust among online shoppers; even quality-conscious students may report reduced trust when production dates are overlooked. Thus, trust is shaped by online shopping, design, recognizability, and economic factors. Feature importance values for all models appear in Figure 6. The results align with decision tree nodes, demonstrating consistency across model visuals. Our findings provide several conclusions and recommendations. Figure 6's Feature Importance Plots show that, as in Ebrahim's thesis, BP is shaped by a range of factors across 22 variables (Ebrahim, 2013). In contrast, BL and BT models emphasize brand awareness, linking it strongly to loy-

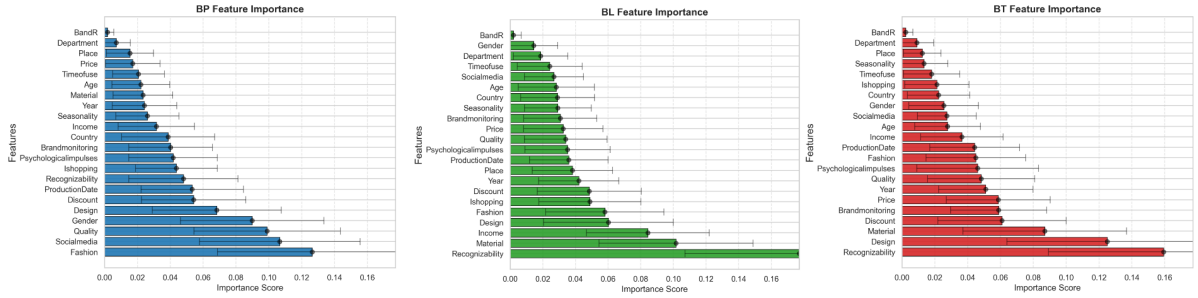

Figure 6. Feature Importance Plots in Balanced BP, BL, and BT Models

alty and trust (Oliver, 2010; Chen, 2021). Feature importance analysis confirms social media's role in shaping brand perception and enhancing marketing impact among students (Pamuksuz et al., 2021). Technological innovations and gender-specific campaigns align with consumer preferences, while materialistic values notably impact BL and BT. Additionally, design, fashion, and social media blend functionality with aesthetics, fostering confidence and proactive loyalty (Oliver, 2010).

RESULTS AND DİSCUSSİON

Our findings align with existing literature on BP, identifying strategic factors influencing BL and BT among young consumers, such as online shopping, social media engagement, student demographics, product quality, and innovation (Zehir et al., 2011; Gumus, 2016; Bastos and Levy, 2012). Companies targeting students are advised to adopt personalized strategies using social media algorithms and AI tools, particularly AI-driven recommendation systems that enhance trust in online shopping (Chawla et al., 2002; Yang et al., 2015). These insights reinforce the quality-loyalty relationship highlighted by Zehir et al. (2011) and Gumus (2016), with Bastos and Levy (2012) noting the growing complexity of brands, making accessible and diverse online options crucial. Yang et al. (2015) further stress the importance of integrating personality traits into BP analysis through social media data and psychometric surveys. Methodologically, SMOTE effectively addresses class imbalance (Chawla et al., 2002), though binary classification and alternative resampling methods may be beneficial in cases of severe imbalance. RF algorithms are effective in determining feature importance, and combining them with coefficient-based models, such as logistic regression, improves interpretability (Ebrahim, 2013). Probabilistic methods, including Bayesian Inference and Maximum Likelihood Estimation (MLE), also enhance performance in imbalanced datasets (Bastos and Levy, 2012). In the original dataset, accuracy values for BP, BL, and BT were 0.76, 0.75, and 0.79, respectively, increasing by 0.09, 0.07, and 0.01 post-balancing. Future studies should expand sample diversity by including students from foundation universities and explore techniques like ADASYN, Tomek Links, and other resampling methods for low sample sizes (Yang et al., 2015). Ensemble methods with class weights and multi-class classifiers are recommended for additional accuracy improvements. This study builds on our previous research (Yılmaz et al., 2024), demonstrating the effectiveness of the Random Forest algorithm in multi-class classification. Consistent with Ebrahim (2013), our findings confirm that BP is shaped by multiple factors, establishing a foundation for scalable marketing solutions and suggesting that similar future studies are likely to yield consistent results.

References

Aaker, D. A. (1996). *Managing Brand Equity*. New York: The Free Press.

Alhaddad, A. (2015). Perceived quality, brand image, and brand trust as determinants of brand loyalty. *Journal of Research in Business and Management, 3*(4), 1–8.

Bastos, W., & Levy, S. J. (2012). A history of the concept of branding: Practice and theory. *Journal of Historical Research in Marketing, 4*(3), 347–368.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*, 123–140.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324.

Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees* (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315139470.

Chaudhary, A., Kolhe, S., & Kamal, R. (2016). An improved random forest classifier for multi-class classification. *Information Processing in Agriculture, 3*(4), 215–222.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

Davis, S. M. (1964). *Brand asset management: Profitable growth through your brands* (2nd ed.). San Francisco: Jossey-Bass.

Dong, Y. (2023). Application of user preference mining algorithms based on data mining and social behavior in brand building.

Ebrahim, R. S. (2013). *A study of brand preference: An experiential view* (Doctoral dissertation, Brunel University Brunel Business School).

Gumus, I. (2016). Brand gender, brand personality, and brand loyalty relationship. *Communication in Mathematical Modeling and Applications, 1*(2), 8–41.

Keller, K. L. (2001). Building customer-based brand equity. *Marketing Management, 10*(2), 14–19.

Kursa, M. B. (2014). Robustness of random forest-based gene selection methods. *BMC Bioinformatics, 15*, 1–8.

Lee, J. L., James, J. D., & Kim, Y. K. (2014). A reconceptualization of brand image. *International Journal of Business Administration, 5*(4), 1.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics, 5*, 115–133.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. *AI Magazine, 27*(4), 12.

Moorman, C., Zaltman, G., & Deshpande, R. (1992). Relationships between providers and users of market research: The dynamics of trust within and between organizations. *Journal of Marketing Research, 29*(3), 314–328.

Morgan-Thomas, A., & Veloutsou, C. (2013). Beyond technology acceptance: Brand relationships and online brand experience. *Journal of Business Research, 66*(1), 21–27.

Oliver, R. L. (2010). Consumer brand loyalty. In *Wiley international encyclopedia of marketing*. Chichester, West Sussex, UK: Wiley-Blackwell.

Pamuksuz, U., Yun, J. T., & Humphreys, A. (2021). A brand-new look at you: Predicting brand personality in social media networks with machine learning. *Journal of Interactive Marketing, 56*, 1–15.

Pears, R., Finlay, J., & Connor, A. M. (2014). Synthetic minority over-sampling technique (SMOTE) for predicting software build outcomes. *arXiv preprint arXiv:1407.2330*.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*, 81–106.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development, 3*(3), 210–229.

Saylı, A., Ozturk, I., & Ustunel, M. (2016). Brand loyalty analysis system using K-means algorithm. *Journal of Engineering Technology and Applied Sciences, 1*(3), 107–126.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*(3),

379–423.

Tesfahun, A., & Bhaskari, D. L. (2013). Intrusion detection using random forests classifier with SMOTE and feature reduction. In *2013 International Conference on Cloud & Ubiquitous Computing & Emerging Technologies* (pp. 127–132). IEEE.

Tian, J., Liu, L., Zhang, F., Ai, Y., Wang, R., & Fei, C. (2019). Multi-domain entropy-random forest method for the fusion diagnosis of inter-shaft bearing faults with acoustic emission signals. *Entropy, 22*(1), 57.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460. https://doi.org/-10.1093/mind/LIX.236.433.

Schee, B. A. (2010). Students as consumers: Programming for brand loyalty. *Services Marketing Quarterly, 32*(1), 32–43.

Yılmaz, M. İ., Yıldız, Ö., & Gökmen, Ş. (2024). Marka tercihine etki eden faktörler: Ankara Hacı Bayram Veli Üniversitesi ekonometri ve iktisat bölümü lisans programı öğrencileri üzerine bir uygulama. In *UYIK-2024 Proceedings Book*.

Yang, C., Pan, S., Mahmud, J., Yang, H., & Srinivasan, P. (2015). Using personal traits for brand preference prediction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Acknowledgements

We thank TÜBİTAK BİDEB, the Turkish Statistical Association, IDSSC'24, Ankara Hacı Bayram Veli University, and Sezin Yılmaz for their support in data acquisition and collection. Our gratitude extends to all contributors to this study.

Conflict of Interest

The authors declare that there is no conflict of interest.