

# Karar Ağacına Dayalı Yöntemler

Bölüm

7

**Mert DEMİRSÖZ<sup>1</sup>, Zeynep ÖZEL<sup>2</sup>, Dr. Öğr. Üyesi Harun YONAR<sup>3</sup>,  
Dr. Öğr. Üyesi Aynur YONAR<sup>4</sup>, Doç. Dr. Mustafa Agah TEKİNDAL<sup>5</sup>**

<sup>1</sup> Selçuk Üniversitesi Veteriner Fakültesi Biyoistatistik AD., (Yüksek Lisans Öğrencisi)

<sup>2</sup> Selçuk Üniversitesi Veteriner Fakültesi Biyoistatistik AD., (Yüksek Lisans Öğrencisi)

<sup>3</sup> Selçuk Üniversitesi Veteriner Fakültesi Biyoistatistik AD., harunyonar@hotmail.com

<sup>4</sup> Selçuk Üniversitesi Fen Fakültesi İstatistik Bölümü aynursahin@selcuk.edu.tr

<sup>5</sup> İzmir Katip Çelebi Üniversitesi Tıp Fakültesi Biyoistatistik AD., matekindal@gmail.com

## Bu Ünite Neler Öğreneceksiniz?

- ◆ Bu üniteyi tamamladıktan sonra;
- ◆ Karar ağaçlarının tarihsel sürecini açıklayabilecek,
- ◆ Karar ağaçlarını kavramsal olarak tanımlayabilecek,
- ◆ Karar ağaçları, budama prensiplerini kavrayabilecek,
- ◆ Karar ağaçları yöntemleri uygun biçimde kullanılabilecek,
- ◆ Karar ağaçlarının test ve eğitim seti uygun yöntemlerle en iyi performans verebilecek şekilde belirlenebilecek,
- ◆ Karar ağaçlarının modelinin uygunluğu test edilebilecektir.

## Hedefler

- ◆ Karar ağaçlarının genel yapısını tanımlamak
- ◆ Karar ağaçlarını en uygun şekilde kullanabilmek
- ◆ Karar ağacı yöntemleri ile en iyi performansı elde edebilmek.





lerde önceki modellerden elde edilen hataların çözüldüğü bir tekniktir. Boosting yönteminde regresyon modelinin kestiricisi, eğitim setinin farklı alt örneklemeleri üzerinde eğitilir. İlk olarak yöntem yerine koyularak orijinal veri setinden çekilen  $N$  boyutlu örneklemeler ile eğitilir. Uygulanan bu işlem ile örneklemelerin hangilerinin daha fazla hataya sahip oldukları belirlenir. Hatadaki fazla kavramı, regresyon makine modelleri için kestirim değerleri ile gözlenen değerler arasındaki farkı en fazla olan örneklemeler için kullanılmaktadır. Bu örneklemelerden hatası en fazla olan, ikinci makine modelinde örneklenme ihtimalleri daha çok olacak şekilde eğitim setine tanımlanmaktadır. Yani  $n$ . adımdaki gözlem değerleri  $n-1$ . adımdaki elde edilen sonuçlara bağlıdır. Kestirim değerleri ise ağırlıklı ortanca kullanılarak birleştirilmektedir. Yani kestirim değerlerinde güvenilir olma durumu yüksek olan kestiriciler daha büyük olacak şekilde ağırlıklandırılırlar. Ağırlıklandırılmalar sonucu model elde edilir. Elde edilen model performans metrikleri ile değerlendirme yapılır (James et al. 2013: Duffy et al. 2002: Alpaydın, 2020)

## KAYNAKLAR

1. Alpaydın, E. (2020). Introduction to machine learning. MIT press.
2. Breiman L, Friedman JH, Olshen RA ve Stone CI (1984). Classification and regression trees. Belmont, Calif.: Wadsworth.
3. Biggs D, De Ville B ve Suen E (1991). A method of choosing multiway partitions for classification and decision trees. Journal of applied statistics, 18(1), 49-62.
4. Çalış A, Kayapınar S ve Çetinyokuş T (2014). Veri Madenciliğinde Karar Ağacı Algoritmaları ile Bilgisayar ve İnternet Güvenliği Üzerine Bir Uygulama. Endüstri Mühendisliği, 25(3), 2-19.
5. Duffy, N., & Helmbold, D. (2002). Boosting methods for regression. Machine Learning, 47(2), 153-200.
6. Friedman JH (1991). Multivariate adaptive regression splines. The annals of statistics, 1-67.
7. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. Springer New York
8. Jensen, JR 1996. Introductory Digital Image Processing, Remote Sensing Perspective. Prentice Hall, Usa
9. Kass GV (1980). An exploratory technique for investigating large quantities of categorical data. Journal of the Royal Statistical Society: Series C (Applied Statistics), 29(2), 119-127.
10. Kavzoğlu T ve Çölkesen İ (2010). Karar ağaçları ile uydu görüntülerinin sınıflandırılması: Kocaeli örneği. Harita Teknolojileri Elektronik Dergisi, 2(1), 36-45.
11. Kirchner K, Tölle KH ve Krieter J (2004). The Analysis Of Simulated Sow Herd Datasets Using Decision Tree Technique. Computers and Electronics In Agriculture, 2(42), 111-127.
12. Loh WY ve Shih YS (1997). Split selection methods for classification trees. Statistica sinica, 815-840.
13. Han, J., Kamber M. (2006). Data Mining Concepts and Techniques Details of Book: Data Mining. Concepts And Techniques. USA: Elsevier.



14. Hanley JA (1989). Receiver operating characteristic (ROC) methodology: the state of the art. *Crit Rev Diagn Imaging*, 29(3), 307-335.
15. Hastie, T., Tibshirani, R. ve Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Newyork: Springer.
16. Mehta M, Agarwal R ve Rissanen J (1996). SLIQ: A fast scalable classifier for data mining. *Proceedings of the International conference on Extending Database Technology, France, March*
17. Mingers J (1989). An empirical comparison of pruning methods for decision tree induction. *Machine Learning*, 4, 227-243
18. Molinaro, A.M., Simon, R. ve Pfeiffer R.M. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21(15), 3301-3307.
19. Omary, Z., & Mtenzi, F. (2010). Machine learning approach to identifying the dataset threshold for the performance estimators in supervised learning. *International Journal for Infonomics (IJI)*, 3(3), 314-325.
20. Quinlan JR (1986). Induction of decision trees. *Machine learning*, 1(1), 81-106.
21. Quinlan JR (1992). *C4.5 Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.
22. Quinlan JR (2004). Data mining tools See5 and C5. 0. [http://www. Rulequest. Com/see5-info. Html](http://www.Rulequest.Com/see5-info.Html).
23. Radcliffe, N. J.; and Surry, P. D. (1999). Differential response analysis: Modelling true response by isolating the effect of a single action, in *Proceedings of Credit Scoring and Credit Control VI*. Credit Research Centre, University of Edinburgh Management School
24. Safavian SR, Landgrebe D, (1991). A survey of decision tree classifier methodology. *IEEE Transactions on Systems Man and Cybernetics*, 21, 660-674
25. Santra AK ve Christy CJ (2012). Genetic algorithm and confusion matrix for document clustering. *International Journal of Computer Science Issues (IJCSI)*, 9(1), 322.
26. Shafer JC, Agarwal R ve Mehta M (1996) SPRINT: A scalable parallel classifier for data mining. *Proceedings of the 24th International Conference on Very lрге Databases*, Mumbai, India September.
27. Nacar EN ve Erdebilli B (2021) Makine Öğrenmesi Algoritmaları ile Satış Tahmini. *Endüstri Mühendisliği*, 32(2), 307-320.
28. Fortmann RS (2021) Understanding the bias-variance tradeoff, URL <http://scott.fortmann-roe.com/docs/BiasVariance.html>.
29. Karakuş BA (2018). Derin öğrenme ve büyük veri yaklaşımları ile metin analizi, Doktora Tezi, Fırat Üniversitesi