Sosyal Bilimler Enstitüsü

T.C.
MARMARA ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
YÖNEYLEM ARAŞTIRMASI BİLİM DALI

İSTATİSTİKSEL YÖNTEMLER İLE ÇALIŞANIN İŞTEN AYRILMASININ TAHMİNLENMESİ VE ÜCRET BELİRLENMESİ İLE İŞTEN AYRILMANIN OPTİMİZASYONU

Doktora Tezi

KADİR YILDIZ

İSTANBUL, 2023

Sosyal Bilimler Enstitüsü

T.C.
MARMARA ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
YÖNEYLEM ARAŞTIRMASI BİLİM DALI

İSTATİSTİKSEL YÖNTEMLER İLE ÇALIŞANIN İŞTEN AYRILMASININ TAHMİNLENMESİ VE ÜCRET BELİRLENMESİ İLE İŞTEN AYRILMANIN OPTİMİZASYONU

Doktora Tezi

KADİR YILDIZ

DANIŞMAN: PROF. DR. SAİT ERDAL DİNÇER

İSTANBUL, 2023

T.C.
MARMARA ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ

TEZ ONAY BELGESİ

..... Anabilim Dalı ……..... Bilim Dalı Yüksek Lisans/Doktora öğrencisi ....'nın …… …… …… …… … ….. …… …… …… …… …… ……. …….……. ..... adlı tez çalışması, Enstitümüz Yönetim Kurulunun ..... tarih ve ..... sayılı kararıyla oluşturulan jüri tarafından oy birliği / oy çokluğu ile Yüksek Lisans/Doktora Tezi olarak kabul edilmiştir.

Tez Savunma Tarihi ...../...../.....

Öğretim Üyesi Adı Soyadı			**İmzası**
1	Tez Danışmanı		
2	Jüri Üyesi		
3	Jüri Üyesi		
4	Jüri Üyesi		
5	Jüri Üyesi		

ÖZET

İSTATİSTİKSEL YÖNTEMLER İLE ÇALIŞANIN İŞTEN AYRILMASININ TAHMİNLENMESİ VE ÜCRET BELİRLENMESİ İLE İŞTEN AYRILMANIN OPTİMİZASYONU

YILDIZ, Kadir

Doktora Tezi

Ekonometri ABD

Yöneylem Araştırması Bilim Dalı

Tez Danışmanı: Prof. Sait Erdal DİNÇER

Ocak 2022, 183 sayfa

Dijitalleşme ve otomasyon ile şirketler operasyonel süreçlerini kısaltmak, kayıplarını minimize etmek, kazançlarını maksimize etmek gibi amaçlarla veri toplamak, işlemek ve kullanmaktadırlar. Bankacılık, finans, telekomünikasyon, e-ticaret gibi sektörlerin öncülük ettiği bu veri vizyonu zamanla diğer şirketler de benimsemiş ve çeşitli amaçlarla veriden faydalanmaya başlamışlardır. Verinin kullanıldığı sektörlerin yanı sıra kullanım şekli de teknolojinin gelişmesine paralel olarak değişiklik göstermiştir. Önceleri geçmiş verinin reaktif analizlerle incelenmesi ve raporlama ihtiyacı giderilirken bu çalışmalar yerini zamanla proaktif analizlere, tahminleme ve optimizasyon çalışmalarına bırakmıştır.

İnsan kaynakları alanı başlı başına bir uzmanlık alanı olup şirketler içerisinde işe alım, ücretlendirme, eğitim ve gelişim, performans yönetimi, vb. alanlarında yönetsel ve operasyonel faaliyetleri üstlenmektedir. İnsan kaynakları, doğası gereği analitik ve veri çalışmalarını geriden takip eden birimlerden biri olup son yıllarda bu alanda da çalışmalar yapılmaya başlandığı gözlemlenmiştir. Bu çalışmada, özellikle insan kaynakları alanının seçilme sebebi analitik çalışmaların kullanımının nispeten az olduğu bir sektörde uygulama yapabilme isteğidir. Bu çalışmada, küresel ve çok uluslu bir şirketin halka açık verisi üzerinden

çalışanların işten ayrılmalarının proaktif şekilde tahminlenmesi sağlanmıştır. Bu kapsamda çeşitli makine öğrenmesi modelleri geliştirilerek model performans karşılaştırması sonucunda en uygun model optimizasyon modeline girdi olarak kullanılmıştır. İşten ayrılma ihtimallerinin bulunmasının ardından belirli bütçe kısıtları dahilinde yüksek performanslı çalışanların işten ayrılmasını minimize edecek bir optimizasyon modeli kurulmuştur. Bu çalışma göstermiştir ki ücretlendirme özellikle talebin de yoğun olduğu çalışanlarda işten ayrılmayı tetikleyen en önemli parametrelerden biridir ve ücretlendirmenin doğru yapılması ile çalışanların bir kısmının kazanımı sağlanabilir. Optimizasyon çalışması sonrasında duyarlılık analizi ile ayrıca farklı bütçe kısıtlarında her çalışan için en uygun maaş seviyesi belirlenmiştir.

Anahtar Kelimeler: İnsan Kaynakları, Optimizasyon, Analitik, Makine Öğrenmesi

ABSTRACT

PREDICTION OF EMPLOYEE CHURN WITH STATISTICAL METHODS AND CHURN OPTIMIZATION WITH COMPENSATION PLANNING

YILDIZ, Kadir

Ph.D. Thesis

Econometrics Department

Operations Research Programme

Adviser of Thesis: Prof. Sait Erdal DİNÇER

January 2022, 183 pages

Thanks to digitalization and automation, companies collect, process and use data to improve operational processes, minimize losses and maximize profits. This data vision, pioneered by banking, finance, telecommunications and e-commerce has been adopted by other companies over time to benefit from data for various reasons. In addition to the sectors where data is utilized, the way of use has also changed in parallel with the advancements in the technology. Traditional use of data for reactive analyses and reporting has been replaced by predictive analytics and optimization studies over time.

Human resources is a specific area of expertise and undertakes numerous managerial and operational activities. HR department is one of the departments that lag behind analytical and data studies and has been observed that studies have started in this field in recent years as well. HR field has been selected with the desire to practice in a sector with relatively weak analytics usage. In this study, proactive estimation of employee churn is provided through a public and multinational data. In this context, various ML models were developed, and the best model was inputted in the optimization model. After revealing churn probabilities, an optimization model was established to minimize the turnover of high-performing employees within certain company budget constraints. Study has shown that compensation is one of the key parameters triggering employee churn, especially in skilled employees, and some employees can be recovered with correct compensation. Sensitivity analysis has also been performed to determine the optimal salaries for each employee.

Keywords: Human Resources, Optimization, Analytics, Machine Learning

ÖNSÖZ

Doktora öğrenimim boyunca desteklerini esirgemeyen, yol gösteren, hızlandıran, kaynak yaratan tez danışmanım Prof. Sait Erdal Dinçer'e; tez jürimde yer alarak beni destekleyen, dersleri ve kitapları ile yol gösteren Prof. Tuncay Can'a ve doktora programının en başından itibaren her başım sıkıştığında kapısını çaldığım ve tez jürimde yer alması ile ayrıca mutlu olduğum Dr. Öğretim Üyesi Naciye Tuba Yılmaz Soydan'a tüm destekleri için teşekkürlerimi sunarım. Bunun yanında, doktora sürecini en iyi şekilde yönetmem konusunda destekleri ve sabırları için anneme, babama ve kardeşime; tüm doktora sürecinde işten izin alarak okulumla ilgilenmem konusunda teşvik eden şirket yöneticilerime ve iş arkadaşlarıma; yeri geldiğinde okul ve iş sebebiyle ihmal etmiş olsam da her zaman yanımda olduklarını bildiğim Yasemin Kalafatoğlu, Özlem Turan, Nazmiye Gizem Bacaksızlar Turbic, Pelin Ekmen, Poyraz Metekol, Berrak Kabaalioğlu ve Aslı Acıgöl'e son olarak doktora sürecinde beraber yürüdüğümüz, birbirimizi izleme sınavları sırasında kapıda beklediğimiz arkadaşlarım Şule Güneş Sarıçam, Fatma Avşar ve Serdar Tuğrul'a teşekkür ederim.

Kadir YILDIZ

İSTANBUL, 2023

İÇİNDEKİLER

ÖZET 4
ABSTRACT 6
ÖNSÖZ 7
İÇİNDEKİLER 8
TABLO LİSTESİ 11
ŞEKİL LİSTESİ 13
KISALTMALAR 16
GİRİŞ 17
1. İNSAN KAYNAKLARI YÖNETİMİ 19
1.1. İnsan Kaynakları Yönetimi Tanımı 19
1.2. İnsan Kaynakları Sorumluluk Alanları 19
2. VERİ ANALİTİĞİ 21
2.1. Veri ve Büyük Veri 21
2.1.1. Veri 21
2.1.2. Büyük Veri 21
2.2. Veri Bilimi, Analitiği, Analizi ve Madenciliği 26
3.2.1. Veri Çalışmaları için İsimlendirmeler 26
2.2.1. Veri Analitiği Uygulama Alanları 29
2.2.2. Veri Analitiğinin Önemi ve Duyulan İhtiyaç 30
2.2.3. Veri Analitiğinin Zorlukları 31
2.3. Veri Analitiği Proje Süreçleri 32
2.3.1. Veri Hazırlığı 33
2.3.2. Veri Ön İşlemleri 33
2.3.3. Sonuçlandırma 40
2.4. Modelleme Algoritmaları 42
2.4.1. Gözetimli Makine Öğrenmesi 42
2.4.2. Gözetimsiz Makine Öğrenmesi 55
2.5. Çeşitli Makine Öğrenmesi Konseptleri 56
2.5.1. Hiperparametre Optimizasyonu 56
2.5.2. AutoML: Otomatik Makine Öğrenmesi 58
2.5.3. Açıklanabilirlik 58

2.5.4. Model Performans Metrikleri 59
3. İNSAN KAYNAKLARI ANALİTİĞİ 62
3.1. İnsan Kaynakları Verisi 62
3.2. İnsan Kaynakları Ham Veri Kaynakları 62
3.2.1. İnsan Kaynakları Performans Metrikleri 63
3.2.2. İnsan Kaynakları Verisinin Sahipliği 66
3.2.3. İnsan Kaynakları Verisi Kullanımında Karşılaşılan Sorunlar 67
3.3. İnsan Kaynakları Analitiği Tarihçesi 67
3.4. İnsan Kaynakları Analitiği Süreci 69
3.5. İnsan Kaynakları Analitiği Olgunluk Endeksi Ölçümü 70
3.6. İnsan Kaynakları Analitiği Sorumluluk Alanları 71
3.6.1. İşgücü Planlama Analitiği 71
3.6.2. Çalışan Bulma Analitiği 72
3.6.3. Çalışan İşe Alım Analitiği 72
3.6.4. İş Başlangıcı, Kültürel Uyum ve Bağlılık Analitiği 73
3.6.5. Çalışanın İşten Ayrılması ve Çalışan Muhafazası Analitiği 73
3.6.6. Çalışan Performans Yönetimi ve Ömürboyu Değer Hesabı Analitiği 74
3.6.7. Çalışan Sağlığı ve Güvenliği Analitiği 75
4. ÇALIŞAN İŞTEN AYRILMASI 77
4.1. Konseptler 77
4.2. Hesaplamalar 77
4.3. Çalışan İşten Ayrılmasının En Önemli Sebepleri 78
4.3.1. Demografik Özellikler 81
4.3.2. İş Memnuniyeti 81
4.3.3. Maddi Sebepler 82
4.3.4. Çalışma Koşulları 83
4.3.5. Şirket – Çalışan Uyumsuzluğu 83
4.3.6. Uygun Olmayan İşe Alım Süreci 84
5. ÇALIŞAN KAYBI ANALİTİĞİ VE OPTİMİZASYONU UYGULAMASI 87
5.1. Veri Kümesi 87
5.2. Veri Hazırlığı ve Ön İşlemleri 88
5.3. Veri Analizi 88
5.4. Modelleme 107
5.4.1. Karar Ağacı Model 107

5.4.2. Karar Ağacı Modeli – Hiperparametre Optimizasyonu ile 108
5.4.3. Rassal Ağaç Modeli 109
5.4.4. Rassal Ağaç Modeli - Hiperparametre Optimizasyonu ile 110
5.4.5. Lojistik Regresyon Modeli 111
5.4.6. Destek Vektör Makinesi Modeli 112
5.4.7. Ekstrem Gradyan Artırma Modeli 113
5.4.8. Uyarlanabilir Yükseltme Modeli 114
5.4.9. Catboost Modeli 115
5.4.10. Light GBM Modeli 116
5.4.11. Model Sonuçları Karşılaştırma 117
5.5. Duyarlılık Analizi 119
5.6. Optimizasyon 122
6. SONUÇ 126
7. KAYNAKÇA 127
8. EKLER 136
Ek 1: Çalışma Planı 136
Ek 2: Kodlar 136

TABLO LİSTESİ

Sayfa No.

Tablo 1. Literatürde Büyük Veri Tanımları 23
Tablo 2. Veri Madenciliği ve Veri Analizi Karşılaştırması 27
Tablo 3. Veri Analitiği Uygulama Alanları 29
Tablo 4. İnsan Kaynakları Alanındaki Veri Kaynakları 63
Tablo 5. İnsan Kaynakları Etkinlik Metrikleri 64
Tablo 6. İnsan Kaynakları Etkililik Metrikleri 64
Tablo 7. İnsan Kaynakları İş Çıktısı Metrikleri 65
Tablo 8. İnsan Kaynakları Verisinin Sahipliği 66
Tablo 9. Geleneksel ve Analitik İnsan Kaynakları Yönetimi 75
Tablo 10. Literatür Sınıflandırması 84
Tablo 11. Kullanılacak Veri Kümesi 87
Tablo 12. Veri Tipleri 88
Tablo 13. Veri İstatistikleri 92
Tablo 14. İşten Ayrılma Durumuna Göre İstatistikler 101
Tablo 15. Yeni Türetilen Değişkenler 103
Tablo 16. Türetilen Değişkenlerin İşten Ayrılmaya Göre İstatistikleri 105
Tablo 17. İşten Ayrılma Durumu Bazında Türetilmiş Veri Dağılımları 106
Tablo 18. Karar ağacı modeli performans metrikleri 107
Tablo 19. Karar ağacı modeli performans metrikleri (2) 109
Tablo 20. Rassal ağaç modeli performans metrikleri 110
Tablo 21. Rassal Ağaç Modeli Performans Metrikleri (2) 111
Tablo 22. Lojistik Regresyon Modeli Performans Metrikleri 112

Tablo 23. Destek Vektör Modeli Performans Metrikleri 113
Tablo 24. Ekstrem Gradyan Artırma Modeli Performans Metrikleri 113
Tablo 25. Uyarlanabilir Yükseltme Modeli Performans Metrikleri 114
Tablo 26. Catboost Modeli Performans Metrikleri 115
Tablo 27. Light GBM Modeli Performans Metrikleri 116
Tablo 28. Model Performans Karşılaştırması 118
Tablo 29. Lojistik Regresyon Değişkenleri 118
Tablo 30. Algoritma Bazında Maaş Artışına Göre Duyarlılık Analizi 121
Tablo 31. Optimizasyon Modeli Parametreleri 122
Tablo 32. Optimizasyon Modeli Sonucu 124
Tablo 33. Bütçe için Duyarlılık Analizi Sonuçları 125

ŞEKİL LİSTESİ

Sayfa No.

Şekil 1. Büyük Veri Tanımlarının Sınıflandırılması 25
Şekil 2. Veri Madenciliği, Analitiği, Analizi ve Bilimi 29
Şekil 3. Dünyadaki Veri Büyüklüğü 30
Şekil 4. Veri Analitiği Proje Süreçleri (1) 32
Şekil 5. Veri Analitiği Proje Süreçleri (2) 32
Şekil 6. Veri Analitiği Proje Süreçleri (3) 33
Şekil 7. Veri Boyutu Azaltma Teknikleri 39
Şekil 8. Modellerin Canlıya Alınma Süreci 41
Şekil 9. Makine Öğrenmesi Türleri 42
Şekil 10. Regresyon ve Sınıflama Problemleri 43
Şekil 11. İkili ve Çoklu Sınıflama Problemleri 44
Şekil 12. Yapay Sinir Ağları 45
Şekil 13. Karar Ağacı 47
Şekil 14. Naive Bayes Sınıflandırıcısı 48
Şekil 15. Destek Vektör Makineleri 51
Şekil 16. Artırma Algoritmalarının Akışı 53
Şekil 17. Rassal Ağaç Algoritması 54
Şekil 18. Rassal Ağaç Algoritma Akışı 55
Şekil 19. Gözetimsiz Makine Öğrenmesi 56
Şekil 20. AutoML - Otomatik Makine Öğrenmesi 58
Şekil 21. Hata Matrisi 60
Şekil 22. ROC Eğrisi 61

Şekil 23. İnsan Kaynakları Alanında Veri Kullanımı 62
Şekil 24. İnsan kaynakları alanında anket sonuçları 67
Şekil 25. İnsan Kaynakları Analitiği Olgunluk Endeksi 70
Şekil 26. Numerik Veri Grafikleri 93
Şekil 27. Kategorik Veri Grafikleri 94
Şekil 28. Korelasyon Matrisi 96
Şekil 29. Hedef Değişken Analizi 97
Şekil 30. İşten Ayrılma Durumuna Göre Veri Dağılımları 99
Şekil 31. Sayısal Değişkenler ile Hedef Değişken Arası Doğrusal İlişki 100
Şekil 32. İşten Ayrılma Durumu Bazında Dağılımlar 102
Şekil 33. Türetilen Değişkenlerin İşten Ayrılma Verisine Göre Dağılımı 104
Şekil 34. Türetilen Değişkenler ile Hedef Değişken Arasındaki İlişki 104
Şekil 35. Karar ağacı modeli hata matrisi 108
Şekil 36. Karar ağacı modeli değişken önem dereceleri 108
Şekil 37. Karar ağacı modeli hata matrisi (2) 109
Şekil 38. Karar ağacı modeli değişken önem dereceleri (2) 109
Şekil 39. Rassal ağaç modeli hata matrisi 110
Şekil 40. Rassal ağaç modeli değişken önem dereceleri 110
Şekil 41. Rassal ağaç modeli hata matrisi (2) 111
Şekil 42. Rassal ağaç modeli değişken önem dereceleri (2) 111
Şekil 43. Lojistik regresyon modeli hata matrisi 112
Şekil 44. Lojistik regresyon modeli değişken önem dereceleri 112
Şekil 45. Ekstrem Gradyan Artırma modeli hata matrisi 113
Şekil 46. Ekstrem gradyan artırma modeli değişken önem dereceleri 114
Şekil 47. Uyarlanabilir Yükseltme modeli hata matrisi 114

Şekil 48. Uyarlanabilir Yükseltme modeli değişken önem dereceleri 115
Şekil 49. Catboost modeli hata matrisi 116
Şekil 50. Catboost modeli değişken önem dereceleri 116
Şekil 51. Light GBM modeli hata matrisi 117
Şekil 52. Light GBM modeli değişken önem dereceleri 117
Şekil 53. Optimizasyonun Genel Kullanım Akışı 120
Şekil 54. Optimizasyonun Genel Kullanım Akışı (2) 121
Şekil 55. Çalışma Planı 136

KISALTMALAR

AdaBoost: Uyarlanabilir Gradyan Artırma

ANOVA: Varyans Analizi

AUC: Eğri Altında Kalan Alan

AutoML: Otomatik Makine Öğrenmesi

CART: Sınıflama ve Regresyon Ağaçları

CIPD: Yetkili Personel ve Gelişim Enstitüsü

CRM: Müşteri İlişki Yönetimi

FN: Yanlış Negatif

FP: Yanlış Pozitif

İK: İnsan Kaynakları

LDA: Lineer Diskriminant Analizi

Light GBM: Hafif Gradyan Artırma Makineleri

OECD: Ekonomik İşbirliği ve Kalkınma Örgütü

PAMI: İnsan Analitiği Olgunluk Endeksi

SEM: Yapısal Eşitlik Modellemesi

SHAP: Shapley'in Eklenebilir Açıklamaları

SVM: Destek Vektör Makineleri

TN: Doğru Negatif

TP: Doğru Pozitif

XAI: Açıklanabilir yapay zeka

XGBOOST: Ekstrem Gradyan Artırma

GİRİŞ

Son yıllarda verinin öneminin artması ile birlikte iş dünyası, kamu kuruluşları ve akademi bu alana odaklanmaya başlamıştır. İş dünyasında verinin toplanması ve bundan finansal veya operasyonel bir fayda yaratılması amacıyla organizasyon içerisinde bu alanda çalışacak ekipler kurulmuş ve üniversitelerin ilgili bölümlerinden mezun öğrenciler istihdam edilmeye başlanmıştır. Kamu kuruluşları, kişisel verilere sadık kalmak ile birlikte ellerindeki veriyi birbirleri ile paylaşarak sosyal bir fayda yaratmaya çalışmaktadırlar. Üniversiteler ise yıllardır hem bu özel veya kamu kurumlarında görev alacak işgücünü yetiştirmekte hem de yeni tekniklerin geliştirilmesi veya farklı alanlarda uygulanması için araştırmalar yapmaktadırlar. Uzun yıllardır Mühendislik, İstatistik ve Ekonometri bölümü mezunları sektörde bu alanda istihdam edilirken son yıllarda ülkemizdeki çeşitli üniversitelerin veri bilimi veya yapay zeka alanlarında bölümler açması veriye duyulan önemi yansıtan önemli bir göstergedir.

Veri ile ilgili yapılan çalışmalar öncelikle büyük verinin zorunlu veya ticari nedenlerle toplandığı ve saklandığı bankacılık, finans, telekomünikasyon ve e-ticaret sektörlerinde yoğunlaşmıştır. Bu sektörlerdeki çalışmalar da doğrudan yasal veya ticari çıktılar üzerinde odaklanmıştır. Örneğin, bankacılık sektöründeki yasal raporlamalar, müşterinin temerrüt ihtimalini veya temerrüt halindeki kaybını tahmin etmekte kullanılan istatistiki modeller, müşterinin kredibilitesini ölçen modeller, telekomünikasyon sektöründeki cihaz finansmanında kullanılan limit belirleme modelleri bunlardan bazılarıdır.

Son yıllarda verinin öneminin fark edilmesi ile her sektörde ve sektör içerisinde farklı alanlarda veri toplanmaya başlanmış ve bundan katma değerli fayda yaratma arayışına girilmiştir. Bu fayda ise artık doğrudan fayda olmanın yanında dolaylı fayda sağlayacak çalışmalara da odaklanmaya başlamıştır. Bunlar, veri odaklı bakış açısı ile müşteri memnuniyetini artıracak hizmetler sağlayarak müşteri kaybını azaltmak, yeni müşteri kazanımı elde etmek gibi örnekler şeklinde kendini göstermeye başlamıştır. İnsan Kaynakları alanında yapılan veri odaklı çalışmalar da bu akımın sonucudur.

İş piyasası, hem firmalar hem de çalışanlar için dinamik bir ortamdır ve bazı işyerleri yeni kurulurken bazıları kapanmaktadır, bazılarında yeni pozisyonlar ortaya çıkarken bazılarında pozisyonlar ortadan kalkmaktadır. Bunlar çalışanların gönülsüz iş değişikliği yapmasına zemin oluşturmaktadır. Bunun yanında, çalışanlar bir işten diğerine veya tamamen iş dünyasının dışına

özgürce geçiş yapabilmektedir. İş kaynağının bu şekilde yer değiştirmesi "çalışanın işten ayrılması" kavramını da beraberinde getirmektedir. (Dale-Olsen, H., 2016).

Çalışanın işten ayrılması ile birlikte iş görüşmesi sürecinden kaynaklanan maliyetler, seçim ve işe alım süreci boyunca zaman kaybı, yeni çalışanın eğitim ihtiyacından kaynaklı maliyetler, verimlilikte düşüş, hizmette duraksama, ürün kalitesinde değişim ve bunlara bağlı olarak şirket gelirinde düşüş beklenmektedir (Pérez-Campdesuñer, R., vd., 2018). Şirketlerin bu negatif etkilerden korunup buna göre gelecekteki aksiyonlarını belirlemeleri yararına olacaktır. Çalışanlar için işten ayrılma olasılığının tahmin edilip işten ayrılması olası olan çalışanlar için eğer elde tutulmak isteniyorsa yapıcı aksiyonların alınması gerekmektedir.

Bu tez kapsamında çalışanların işten ayrılmasının tahmini için makine öğrenmesi metotları ile bir tahmin modeli kurulacak olup sonrasında çalışanın hayat boyu değeri hesaplanarak çeşitli kısıtlar altında ve hedefler ile işten ayrılmaları minimize etmek amaçlanmaktadır. Dolayısıyla bu yaklaşımı benimseyen şirketlerin finansal ve motivasyonel fayda sağlaması beklenmektedir.

Bu alanda yapılan çalışmalar son yıllarda yoğunlaşmış olup makine öğrenmesi yöntemleri ile işten ayrılma tahmini ise genelde son 2-3 yıl içerisinde ağırlık kazanmıştır. Ancak yapılan çalışmaların hiçbirinde işten ayrılmanın tahmini sonrasında bir optimizasyon bakış açısı dahil edilmemiştir. Bu tez ile, işten ayrılmanın tahmininden optimize edilmesine kadarki sürecin çizilmesi ve kullanıma hazır hale getirilmesi planlanmaktadır.

1. İNSAN KAYNAKLARI YÖNETİMİ

Bu bölümde insan kaynakları yönetimi kavramı tanıtılmış ve sorumluluk alanları özetlenmiştir. İlerleyen bölümlerde bu sorumluluk alanlarındaki analitik uygulamalar detaylı şekilde anlatılmıştır.

1.1. İnsan Kaynakları Yönetimi Tanımı

İnsan kaynakları yönetimi, 1996'de Ulrich'in Harvard Business Review tarafından basılan Human Resource Champions isimli kitabında insan kaynağı ile iş stratejisi arasında uyumun sağlanması, organizasyon süreçlerinin yeniden tasarımı, çalışanların dinlenip cevap verilmesi, transformasyon ve dönüşüm yönetimi olarak tanımlanmıştır. Pratikte ise, insan kaynakları yönetimi çalışanın tüm çalışma yaşam döngüsü içerisindeki tecrübesinden sorumludur. Bunun içerisinde doğru çalışanların şirkete çekilmesi ve işe alım süreci, işe alınan çalışanların eğitim süreci, performans yönetimi süreci, ücretlendirme, iş hukukuna uyum gibi roller barınmaktadır (Obedgiu, V., 2016).

İnsan Kaynakları'nın konsept olarak var olmasından çok daha önceleri bu alanda çalışmalar yapıldığı ve bilinen tarihinin M.Ö. 5 milyon yıllarına dayandığı söylenebilir. Taş Devri'nde ilk avcı toplayıcı ve taştan alet geliştirici olan topluluklarda usta çırak ilişkisi insan kaynaklarının temelini oluşturmaktadır. M.Ö. 3000 yıllarında Mısır Piramitleri'nin geliştirilmesi de zamanın şartlarına göre devasa bir projenin uygun insan kaynakları yönetimi ile nasıl tamamlanabildiğine güzel bir örnektir. Sonrasında Helenik dönem, Orta Çağ, Aydınlanma Çayı, Sanayi Devrimi ve 2. Dünya Savaşı dönemlerinin her biri mevcut bilgi birikimi ve yöntemler üzerine yenilerini eklemiş ve günümüzdeki İnsan Kaynakları alanı doğmuştur. 1980 yılına gelindiğinde Hansen, ilk defa insan kaynakları yöneticisinin sorumluluk alanlarını açıklamıştır. Günümüzde ise bu alanlar hem genişlemekte hem de çalışma şekilleri teknoloji ve veri kullanımı ile birlikte değişerek zamana ayak uydurmaktadır (Gosney, M. W. & Hughes, C., 2016). İnsan kaynakları konseptinin felsefi geçmişi bu şekilde olmakla birlikte analitik yöntemler ile desteklenmesinin detaylı tarihçesi ilerleyen bölümlerde anlatılacaktır.

1.2. İnsan Kaynakları Sorumluluk Alanları

Şirketlerdeki en iyi çalışanlar, en iyi ürün veya hizmetleri üretmekte, en yüksek karlılığı getirmekte, en verimli çalışma yöntemlerini bulmaktadır. Bu çalışanlar ayrıca işletmenin mevcut iş arkadaşları için daha iyi bir yer olmasını, yeni yetenekli iş arkadaşlarının kazandırılmasını ve müşteriler

için iş birliğinden memnun olunmasını sağlamaktadır. Bu üstün yetenekli çalışanları bulmak, istihdam etmek, yönetmek ve şirkette tutmak şirketlerin geleceği için çok önemlidir (Isson & Harriott, 2016).

2014 yılında Harvard Business Review dergisinde yayınlanan ve Deloitte ve Visier firmaları tarafından gerçekleştirilen bir ankette sektörün önde gelen firmalarından 230 tanesinin yöneticisinin görüşleri alınmıştır. Bu şirketlerden %38'i 10 bin üzeri, %50'si ise 5 bin üzeri çalışana sahip şirketlerde çalışmakta olup şirketler Amerika, Avrupa, Asya ve Afrika bölgelerinde dağılmış durumdadır. Bu çalışmada şirketlerin insan kaynakları alanındaki en büyük üç kaygıları ve bunu bildiren yöneticilerin oranı şu şekilde görülmüştür (Harvard Business Review Analytic Services, 2014):

1. Üstün yetenekli çalışanların işe alımı (%48)
2. Üstün yetenekli mevcut çalışanların liderlik rolleri için hazırlanması (%46)
3. Üstün yetenekli mevcut çalışanların şirkette tutulması (%43)

İnsan kaynakları içerisindeki sorumluluk alanları şu şekilde sıralanabilir ve her birinde veriye dayalı çeşitli analitik yöntemlerin uygulanması mümkündür (Isson & Harriott, 2016):

1. İşgücü planlama
2. Çalışan bulma
3. Çalışan işe alımı
4. İş başlangıcı, Kültürel Uyum ve Bağlılık
5. Çalışanın İşten Ayrılması ve Çalışan Muhafazası
6. Çalışan performans değerlendirme, geliştirme ve ömür boyu değer hesabı
7. Çalışan sağlığı ve güvenliği

2. VERİ ANALİTİĞİ

2.1. Veri ve Büyük Veri

Veri analitiği çalışmasının anlaşılabilmesi ve doğru uygulanabilmesi için öncelikle bu bölümde verinin tanımı ve veri analitiği konsepti anlatılmaktadır.

2.1.1. Veri

Veri, "bir konu hakkında araştırma, tartışma, bilgi edinme, akıl yürütme sonucunda oluşmuş olan, işlenmemiş, yorum yapmaya imkân verecek düzeyde sistemleştirilmemiş ham bilgi" şeklinde tanımlanmaktadır. Bu kapsamda, bilgi; geçerli, yeni, yararlı ve insanlar tarafından anlaşılabilir ilginç örüntüler olarak tanımlanmaktadır (Runkler, 2016).

Gahi, Y., v.d. makalelerinde veriyi aşağıda detayları sunulan üç sınıf içerisinde incelemişlerdir (2016):

1. **Yapısal veri:** Modellenmesi, girilmesi, depolanması, sorgulanması, işlenmesi ve görselleştirilmesi kolay veri tipidir. Genellikle daha önceden tipi ve boyutu belirlenmiş olan ve ilişkisel veri tabanı veya tablolarda tutulurlar. Veri işleme yüksek performans veya paralel işleme gerektirmediğinden yapısal veri faydalı bilginin çıkarılmasını daha kolay kılmaktadır.
2. **Yarı yapısal veri:** Rasyonel bir yapıda bulunmayan ama modellemeyi yine de kolaylaştıran bazı özelliklere sahip veri tipidir. Etiket ve işaret gibi belli elemanları tanımlamayı sağlayan bilgiler yer aldığından veri içerisindeki alanların hiyerarşik gösterimi yapılabilmektedir. En çok bilinen yarı yapısal veriler XML (Extensible Markup Language) ve JSON (JavaScript Object Notation) veri dosyalarıdır.
3. **Yapısal olmayan veri:** Belirli olmayan bir formatta gösterilen ve depolanan veri tipidir. Genellikle serbest formatta yazı şeklinde olan kitap, makale, doküman, e-posta veya fotoğraf, ses, video gibi medya dosyaları bu tipe örnek verilebilir. Bu tip veri kolay görselleştirilebilir veya işlenebilir bir formatta değildir ve bu sebeple NOSQL yeni veri işleme yöntemleri kullanılmaktadır.

2.1.2. Büyük Veri

"Büyük veri" kavramı bu anılma biçimi ile ilk kez 1997 yılındaki 8. IEEE Görselleştirme Konferansı'nda Cox ve Ellsworth tarafından "Çekirdek Dışı Görselleştirme için Uygulama Kontrollü

Talep Çağrısı" isimli konferans raporlarında kullanılmıştır. Bu çalışmada, veri kümelerinin büyük olması ile ana bellek, yedek disk ve harici disklerin boyutlarının yetersiz kalması sebebiyle görselleştirmenin bilgisayarlar için zor olduğundan bahsedilirken sorunun "büyük veri" sorunu olduğu dile getirilmiştir (Cox, Ellsworth, 1997) .Cox ve Ellsworth 1999 yılında daha büyük bir araştırmacı grubu ile yazdıkları yeni bir makalede, "Gigabayt veri kümelerinin gerçek zamanlı görsel olarak incelenmesi", yine bu konsept üzerinde çalışmaya devam ettiklerini göstermişlerdir (Bryson, vd., 1999). 2008 yılına kadar sınırlı sayıda yer verilen büyük veri kavramı bu tarihten sonra literatürde daha fazla yer almaya başlamıştır. Proquest Kütüphanesi üzerinde yapılan bir araştırmaya göre 2008 yılında bu terimin literatürde kullanımı çok sınırlı iken 2010'da yaklaşık 2 katına çıkmış, 2011'de bunun 12 katına, 2012'de neredeyse 40 katına çıktığı görülmüştür (Gandomi, Haider, 2015).

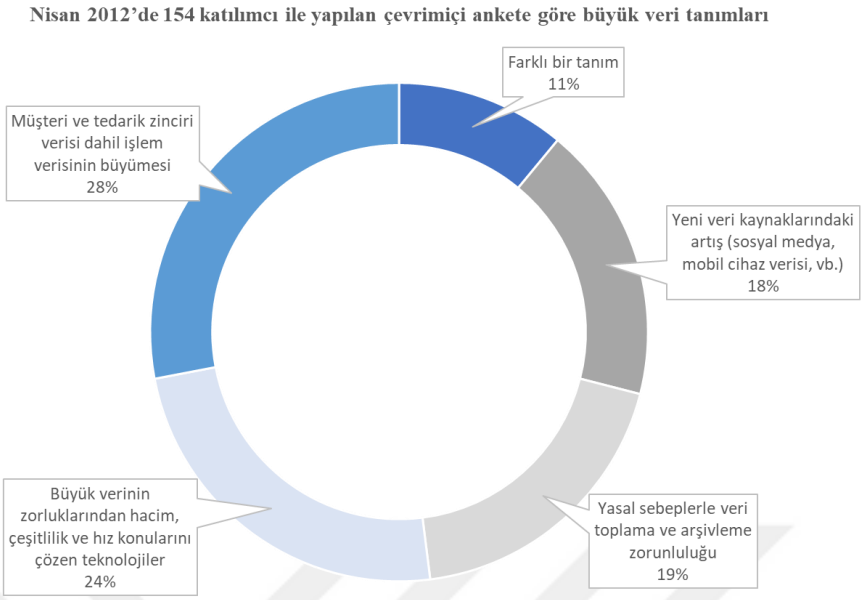
Büyük veri, geleneksel yazılım ve depolama araçları ile tolere edilebilir bir süre içerisinde işlemenin zor olacağı ve esneklik sağlamayacağı büyüklükte verileri tanımlamak için kullanılan bir terimdir. Büyük verinin boyutu düzinelerce terabayttan (TB) petabayt (PB) düzeyine kadar artarak hızla yükselmektedir. Bu sebeple büyük verinin elde edilmesi, depolanması, aranması, paylaşılması, analitiği ve görselleştirilmesi bu alandaki çalışmalarda karşılaşılan en büyük zorluklardandır (Kubick, 2012). Öte yandan Oxford Sözlüğü'nün büyük veri tanımı ise şu şekildedir: "Büyük veri, özellikle insan davranışları ve etkileşimi konularında numerik analizlerle örüntü, trend ve bağlantıların ortaya çıkarılması için kullanılan ekstrem büyüklükteki veri kümeleri". Diğer araştırmacılar bu tanımı eksik bulsalar da büyük verinin, geleneksel veri tanımından farklı şekilde tanımlanması gerektiği konusunda hemfikirdir (Al-4Shiakli, 2019). Bahsedilen bu büyüklükteki veriler ile çalışmak için de daha ileri teknikler ve araçlar kullanılması zorunlu olmuştur.

Bu tanımlamalar dışında da literatürde büyük verinin çeşitli tanımlamaları yapılmıştır. Aşağıdaki tablo bu tanımlamalar ve yorumlamalardan bir kısmını içermektedir ve tanımlamaların ve çalışmaların içinde bulunulan yıl ve bu yıldaki teknolojik ve sosyal gelişmelere bağlı olduğu görülebilmektedir. Örneğin; bilgisayar teknolojisindeki gelişmeler veya Covid-19 salgını gibi.

Tablo 1. Literatürde Büyük Veri Tanımları

Yazar	Yıl	Tanım
Manyika, J. vd.	2011	Boyutu tipik veri kümesi yazılım araçlarının yakalayacağı, depolayacağı, yöneteceği ve analiz edeceği yeteneklerin ötesinde olan veri kümeleridir
Boyd, D., Crawford, K.	2012	Büyük veri, kültürel, teknolojik ve ilmi bir olgudur ve teknoloji, analiz ve mitoloji ile etkileşim içerisindedir.
Fisher, D, vd.	2012	Büyük veri, genellikle standart yöntemlerle ele alınıp işlenemeyen veridir
Corbett, P. vd.	2013	Büyük veri günümüz dünyasında bir iş zorunluluğudur ve sigorta şirketlerinin uzun soluklu iş zorluklarının çözümü için kullanılmaktadır. Büyük veri, süreçlerin, organizasyonların ve sigorta sektörünün dönüştürülmesi için yeni yollar sunmaktadır.
Kemp, R.	2014	Büyük veri, günümüzde rekabet avantajı sağlamak amacıyla hızlıca büyüyen yapısal olan ve olmayan dijital veri kümelerinin toplanması ve analiz edilmesi için geliştirilmiş tekniklerin bir stenografisidir.
Gandomi, A., Haider, M.	2015	Büyük veri dış dünyadan bağımsız değerlendirildiğinde değersizdir ve gerçek değeri karar verme amacıyla kullanıldığı zaman ortaya çıkar. Bu sebeple organizasyonlar kanıtlara dayalı karar vermede yüksek hacimli ve hızlı değişen veriden anlamlı iç görüler sağlamak için büyük veri süreçlerine ihtiyaç duyarlar.
Shin, D.	2016	Büyük veri basit bir teknoloji değil, aksine toplum tarafından yaratılan ve ticari ve kamusal kuruluşlar tarafından toplanan muazzam bir ham bilgiden üretilen bir olgudur. Bu olgu, hem fırsat hem de zorluklar getirmektedir ve çift yönlülük fen, teknoloji ve toplum bilimleri için değişik araştırma konuları yaratmaktadır.

Khan, S., vd.	2017	Dijital dünya bir veri patlaması yaşamaktadır ve düzenli şekilde artan bir veri söz konusudur. Büyük veri hem büyük hacimli, çeşitli ve hızla artan verinin depolanması ve işlenmesini sağlayacak sistem boyutuyla hem de kompleks analitik yöntemler ve algoritmalar boyutuyla bilgisayar ve hesaplama yoğun bir daldır.
Acharya, A., vd.	2018	Büyük veri en moda sözcüklerden biri olmuş ve birçok araştırmaya da konu edilmiştir. Bilgilerin birleştirilmesi ve tüm bilgiler ışığında karar verme noktasında büyük veri kapsayıcı ve organizasyonun tümünü karara dahil eden bir rol üstlenmektedir
Mikalef, P., vd.	2019	Büyük veri analitiği birçokları tarafından inovasyon, rekabet ve üretkenlik için bir sonraki sınır olarak değerlendirilmektedir ve bu nedenle hem akademi hem de çalışanlar tarafından büyük ilgi görmektedir.
Madaan, N., vd.	2020	Büyük veri, ölçeği, dağılımı, çeşitliliği ve/veya zamanlılığı yeni teknik mimarilerin, analitik yöntemlerin ve araçların kullanımını zorunlu kılan ve amacı iş değeri için yeni içgörüler elde etmek olan veridir.
Niu, C., Wang, L.	2021	Üretim sektöründe büyük veri, birçok kaynaktan gelen heterojen veriyi ifade etmektedir ve bu veriler ürünün üretim döngüsü içerisinde “Internet of Things” veya RFID (Radyo Frekansı ile Tanımlama) ile ürün, makine ve teçhizatın gerçek zamanlı izlenmesi sırasında oluşmaktadır.
Latifian, A.	2022	Büyük veri, küresel veri boyutlarında aşırı artış ile birlikte ortaya çıkan Covid-19 verisi gibi muazzam büyüklükteki verileri ifade etmektedir.



Şekil 1. Büyük Veri Tanımlarının Sınıflandırılması

Büyük veri tanımlanırken genellikle büyük verinin boyutlarından söz edilerek de tanımlama yapılmıştır. Laney bir makalesinde büyük verinin V harfi ile başlayan 3 boyutu olduğunu belirtmiş ve bu boyutları şu şekilde sıralamıştır: Volume (Hacim), Variety (Çeşitlilik) ve Velocity (Hız). Bunlar şu şekilde tanımlanmıştır (2001):

Volume (Hacim): Verinin büyüklüğünü ifade etmektedir ve bu boyutlar terabayt veya petabayt şeklinde değişmektedir. IBM'in 2012 yaptığı bir ankete göre 144 katılımcının yarısından fazlası 1 terabayt üzerinde hacme sahip verileri büyük veri olarak adlandırdıklarını dile getirmişlerdir. Verinin büyüklüğü zamana ve verinin türüne göre değişmekle birlikte metin verisi veya görsel veriler yapısal verilere göre daha fazla yer kaplayabilmektedir.

Variety (Çeşitlilik): Verinin yapısal heterojenliği ifade etmektedir. Yapılan bir araştırmaya göre verilerin sadece %5'i yapısal veri şeklindedir (excel raporları veya veri tabanlarında saklanmaktadır). Bunun dışındaki kısmı ise metin, resim, ses, video şeklinde bulunmaktadır. Büyük veri yüksek çeşitlilik içermektedir.

Velocity (Hız): Verinin üretilme sıklığını ifade etmektedir. Sosyal medya, akıllı telefonlar, sensörler, dijitalleşme ile birlikte bankacılık ve e-ticaret gibi alanlarda artık veri üretilme hızı çok artmış durumdadır.

Laney'in veri ile ilgili yukarıda sıraladığı 3V dışında verinin diğer boyutları ise şöyle sıralanabilir (Gandomi, Haider, 2015):

Veracity (Doğruluk): IBM verinin dördüncü boyutu olarak doğruluk özelliğini listelemiştir. Örneğin, müşterinin geri bildirimleri firma için veri özelliği taşımaktadır ancak bunların işlenerek pozitif/negatif duygu şeklinde sınıflandırılması ve doğru şekilde değerlendirilmesi de büyük verinin kapsamına girmektedir.

Variability (Değişkenlik): Verinin akış hızındaki değişkenliği ifade etmektedir. Büyük verinin akışkanlığı genellikle değişkenlik göstermektedir ve periyodik olarak iniş çıkışlar mevcuttur.

Value (Değer): Oracle'ın veri ile ilgili listelediği son özelliğe göre ham veri değer bakımından zayıftır ancak bu veri işlenerek katma değerli bir veri kümesi yaratılabilir.

2.2. Veri Bilimi, Analitiği, Analizi ve Madenciliği

Bu alt bölümde veri analizi ve veri analitiği kavramları açıklanacak ve aralarındaki farklar netleştirilecektir.

3.2.1. Veri Çalışmaları için İsimlendirmeler

Runkler, veri analizi ve adımlarını detaylandırdığı kitabında yaptığı tanımda veri analizini, "verinin alınması, işlenmesi, düzenlenmesi, görselleştirilmesi, analiz edilmesi ve izlenmesi" adımları ile tanımlamaktadır. Veri analizi birçok yöntemi içerisinde barındıran geniş kapsamlı bir kavramdır ve veriden bilgi ortaya çıkarma çabasının ve yöntemlerinin tümünü kapsamaktadır. Ayrıca, endüstride veri analizinin esas amacı süreçlerin optimize edilmesi ve şirketlerin rekabetçi pozisyonunun geliştirilmesi olarak açıklanmıştır (Runkler. 2016).

Veri analizi için sözlük tanımı da Gee tarafından dolandırıcılık alanı üzerine yoğunlaşan ve veri analizi ile buna çözüm bulmaya çalışan kitabında verilmiştir. Buna göre veri analizi, analitik yöntemler ve mantıksal gerekçeler ile veri içerisindeki tüm elemanların değerlendirilme sürecidir. Bu değerlendirme adımı bir araştırma çalışması içerisindeki çok sayıdaki adımdan yalnızca biridir. Birçok kaynaktaki veri toplanmakta, incelenmekte ve belirli bir sonuç elde etmek üzere analiz edilmektedir. Analiz yöntemleri çok çeşitli olmakla birlikte veri görselleştirme, veri madenciliği, metin analizi veya iş zekası bu analiz yöntemlerinden bazılarıdır (Gee, 2015).

Verinin analizi ile ilgili ilk çıkan kavramlardan biri olan veri madenciliği kavramının geçmişi 1980'lere kadar gitmektedir (Runkler. 2016). Veri madenciliği için sözlük tanımı "çok geniş veri

kümelerini elekten geçirerek faydalı bilgiler elde etme işlemi" şeklindedir. Veri madenciliği, yapay zeka teknikleri, yapay sinir ağları ve kümeleme analizi gibi gelişmiş istatistiksel araçlar yoluyla veri içerisindeki eğilimleri, örüntüleri ve ilişkileri ortaya çıkarmaya çalışmaktadır. Bu da herhangi bir hipotez testi olmadan veya herhangi bir spesifik çıktı beklenmeden yapılmaktadır (Gee, 2015).

Veri analizi ve veri madenciliği kavramlarının tanımları karşılaştırıldığında anlaşılmaktadır ki veri madenciliği, veri analizinin bir alt kümesidir. Veri analizi bir hipotezin test edilmesi ile başlar ve doğru veya yanlış bir sonuca ulaşılması hedeflenmektedir. Veri analizi ve veri madenciliği farkı daha açıklayıcı olması için madde madde aşağıdaki tabloda karşılaştırılmaktadır (Geekforgeeks, 2021):

Tablo 2. Veri Madenciliği ve Veri Analizi Karşılaştırması

Kriter	Veri Madenciliği	Veri Analizi
Tanım	Geniş veri kümeleri içerisinden önemli örüntülerin çıkarılması sürecidir	Faydalı bilgiler elde etmek için ham verinin analiz edilmesi ve organize edilmesi sürecidir
Fonksiyon	Ham verideki gizli örüntülerin keşfedilmesi için kullanılır	Veri kümesi ile yapılan tüm operasyonlar bir sonuca bağlama amacı ile yapılır
Veri Kümesi	Genellikle geniş ve yapısal veriler ile çalışılır	Veri kümeleri boyut olarak büyük, orta, küçük olabilir; veri kümeleri yapı olarak yapısal, yarı yapısal veya yapısal olmayan formda olabilir.
Modeller	Matematiksel ve istatistiksel modeller gerektirir	Analitik ve iş zekası modelleri gerektirir
Görselleştirme	Görselleştirme şart değildir	Görselleştirme yapılması gerekir
Hedef	Verinin kullanılabilir hale getirilmesi hedeflenir	Veriye dayalı karar verilmesi hedeflenir

Gerekli Bilgi	Makine öğrenmesi, istatistik ve veri tabanı bilgileri kullanılmaktadır	Bilgisayar bilimleri, istatistik, matematik, yapay zeka ve makine öğrenmesi ile alan uzmanlığı gereklidir
Çıktı	Veri eğilim ve örüntüleri	Kabul edilen veya edilemeyen hipotezler

Veri analizi ve veri madenciliği tanımlandığına göre artık veri analitiği tanımlaması yapılabilir. Veri analitiği veri kümesi hakkında öngörü sağlar, kolayca gözlemlenemeyen veri ilişkileri ve yapılarını keşfetmeyi kolaylaştırır, varsayım ve hipotezlerin test edilmesi olanağını getirir, nedensel ilişkileri ortaya çıkarır ve verideki anomalilerin tespitine yardımcı olur. Veri analitiğinin keşfedici analitik (EDA), doğrulayıcı analitik (CDA) ve kalitatif analitik (QDA) olmak üzere de üç alt sınıfa ayrılabilir. Keşfedici analitik verinin keşfedildiği ve veri alanları arasındaki ilişkilerin henüz tam olarak bilinmediği ilk adımdır. Bu adımda hipotezler oluşturulur ve değişkenler arasındaki örüntüler ortaya çıkarılırken genellikle de görsel ve grafik teknikler kullanılır. Doğrulayıcı analitik adımında ise hipotezlerin doğru veya yanlışlığı test edilirken örneklemden elde edilen sonuçlar tüm veri üzerinde de doğrulanır. Veri içerisindeki bağımlı ve bağımsız değişken arasındaki sebep sonuç ilişkileri bu adımda ortaya çıkarılır. Son olarak, kalitatif analitikte ise sayısal olmayan fotoğraf veya metin veya metin gibi verilerden sonuç çıkarılmaya çalışılır (Gee, 2015).

2000'li yıllarda popüler hale gelen veri analitiği, veriler incelenerek örüntülerin tespit edildiği ve buna göre geleceğin tahminlendiği, tahminlenen geleceğe göre ise aksiyonların belirlendiği bir yönetim konseptidir (Deeba, 2020). Bu yöntemde karar verme amacıyla büyük veri kümelerinin bilgisayar sistemleri yardımıyla analiz edilmesi vurgusu yapılmaktadır (Tyagi, 2003). Ayrıca verilerin toplanması, ön işlemden geçirilmesi, analiz edilmesi ve değerlendirilmesi için de istatistik, makine öğrenmesi, örüntü tanıma, sistem teorisi, yöneylem araştırması veya yapay zeka teknikleri kullanılmaktadır (Runkler, 2016).

Şu ana kadar tanımlanan veri madenciliği, veri analizi ve veri analitiği kavramlarının hepsini kapsayan son kavram ise veri bilimidir ve tüm kavramların birbirinden ayırt edilmesinde aşağıdaki görsel kolaylık sağlayacaktır (Medium, 2022):



Şekil 2. Veri Madenciliği, Analitiği, Analizi ve Bilimi

2.2.1. Veri Analitiği Uygulama Alanları

Literatürde yapılan incelemeler sırasında Veri Analitiği uygulama alanlarının çok çeşitli olduğu gözlemlenmiştir. Finans, sigorta, telekomünikasyon, bilgi teknolojileri, sağlık, gıda, turizm, akıllı şehircilik, yönetim bilimleri büyük veri uygulamalarının en çok görüldüğü alanların başında gelmektedir. Çeşitli uygulama alanlarından güncel referanslar aşağıdaki tabloda listelenmiştir:

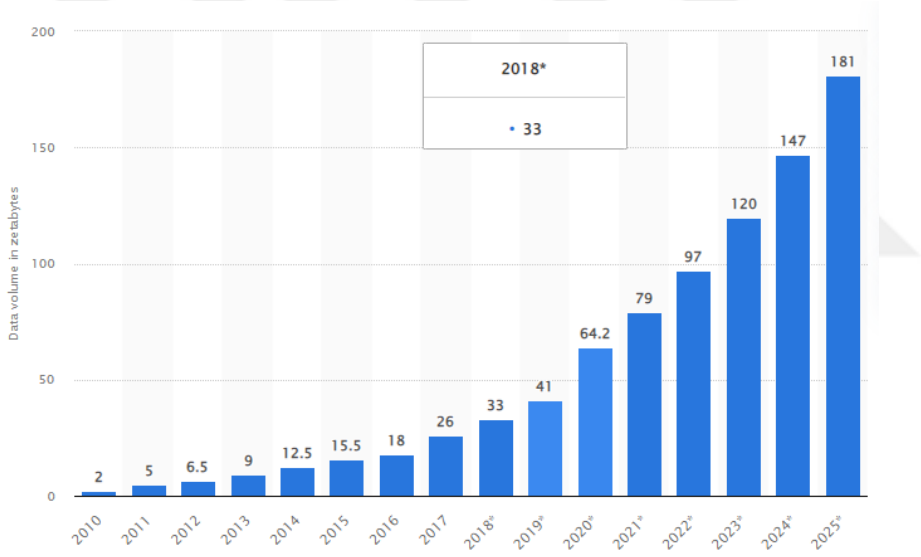
Tablo 3. Veri Analitiği Uygulama Alanları

Uygulama Alanı	**Yazarlar**
Finans	Subrahmanyam, A. (2019)
Firma yönetimi	Mikalef, P., v.d. (2019)
Güvenlik	Yılmaz, G. (2019)
Havacılık	Hazırbaba, N. (2020)
Hukuk	Kemp, R. (2014)
Moda/Alışveriş	Işik, H. (2020)

Müşteri hizmetleri	Güçlükol, S. (2020)
Sağlık	Yang, Y., vd. (2021)
Siber sistemler	Niu, C., & Wang, L. (2022)
Sigortacılık	Corbett, P., Schroeck, M., & Shockley, R. (2013)
Telekomünikasyon	He, Y., v.d. (2016)

2.2.2. Veri Analitiğinin Önemi ve Duyulan İhtiyaç

Dünyada oluşturulan, kopyalanan ve tüketilen veri hızla büyümektedir ve 2020'de 64.2 zetabayt seviyesine kadar ulaşmıştır. Hatta 2025'e kadar ise 180 zetabayt seviyelerine kadar artması beklenmektedir. Buradaki büyüme eğrisi, Covid-19 pandemi etkisi ile birlikte daha da dikleşecek şekilde güncellenmiştir (Statista, 2021).



Şekil 3. Dünyadaki Veri Büyüklüğü

Bilinen veri kaynaklarındaki verilerin her gün yeni eklemeler ile büyümesinin yanı sıra daha önce popüler olmayan veya hiç var olmayan yeni veri kaynaklarının ortaya çıkması da bu büyümeyi ve veri analitiği ihtiyacını tetiklemektedir. Pandemi öncesinde e-ticaret sektöründe biriken veri ile 2022 Türkiye'sinde biriken veri arasında çok büyük fark vardır veya bitcoin dünyasındaki veri hayatımıza çok yeni girmiştir, Metaverse verisi de yeni yeni hayatımıza girmek üzeredir. Bunlar düşünüldüğünde yeni veri analitik yöntemleri hayatımıza girmeye devam edecek ve bunlar kullanılarak günlük hayat problemlerine daha iyi çözümler bulmak zorunlu hale gelecektir. İnsan kaynakları da uzun yıllardır var

olan bir konsept olmasına rağmen yeni çalışan davranışları, yeni veri türleri ve bu başlık altındaki yeni alt problemler ile daha da gelişecektir.

2.2.3. Veri Analitiğinin Zorlukları

Bu bölüm büyük veri ile analitik çalışmalar yapmanın zorluklarını listelemektedir (Assunçao, M., v.d., 2015, Al-Shiakli, S. 2019).

Veri çeşitliliği: Sürekli boyutu artan bir verinin yönetimi zordur. Özellikle de yapısal olmayan veriden anlamlı bir fayda çıkarmak veya farklı kaynaklardan verilerin bir araya getirilmesini sağlamak en büyük zorluklardır.

Veri saklama: Özellikle yapısal olmayan veri içerisinde olmak üzere verimli bir şekilde anlamlı bilgileri tespit etmek ve saklamak zordur. Saklama konusundaki ihtiyaç ise büyük boyutlu verileri hızlıca geri getirilebilir ve ulaşılabilir şekilde depolamaktır.

Veri entegrasyonu: Farklı tür ve kaynaklardaki verileri yönetebilen yeni protokoller ve arayüzler ile verinin entegrasyonu bir başka zorluktur.

Veri işleme ve kaynak yönetimi: Akan veya çok boyutlu veriler için optimal modeller, en iyi dosya sistemlerini yönetebilen uç sistemler, farklı programlama modellerinden uygulamaları bir araya getiren cihazlar, vb.

Veri güvenliği: Birçok farklı kaynaktan toplanan çeşitli format ve boyutlardaki verilerin bir araya getirilmesi sırasında güvenlik açıkları olmamalıdır. Herhangi bir güvenlik sorunu firmaların finansal kayıplara uğramasına neden olabilir ve bunun önüne geçilmesi için gerekli önlemler alınmalıdır.

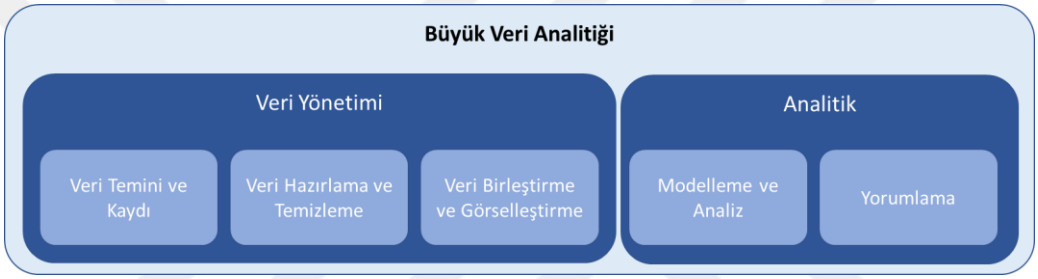
Veri gizliliği: Verilerin yasal veya prosedürel sebeplerle bazen modifiye edilmesi veya maskelenmesi gerekebilmektedir. Bazı kullanım alanlarında bu işlemin nasıl yapılacağına kadar tüm detaylar yönetmelikler ile belirlenmiştir.

Veri kalitesi: Veri kaynaklarından alınan ham veride, birleştirme esnasındaki operasyonlardan kaynaklı olarak nihai veride veya dış etkenler sebebi ile veri kalitesinde bozulmalar olabilmektedir. Bunun sürekli kontrol edilerek gerekli düzenlemelerin yapılması sağlanmalıdır.

2.3. Veri Analitiği Proje Süreçleri

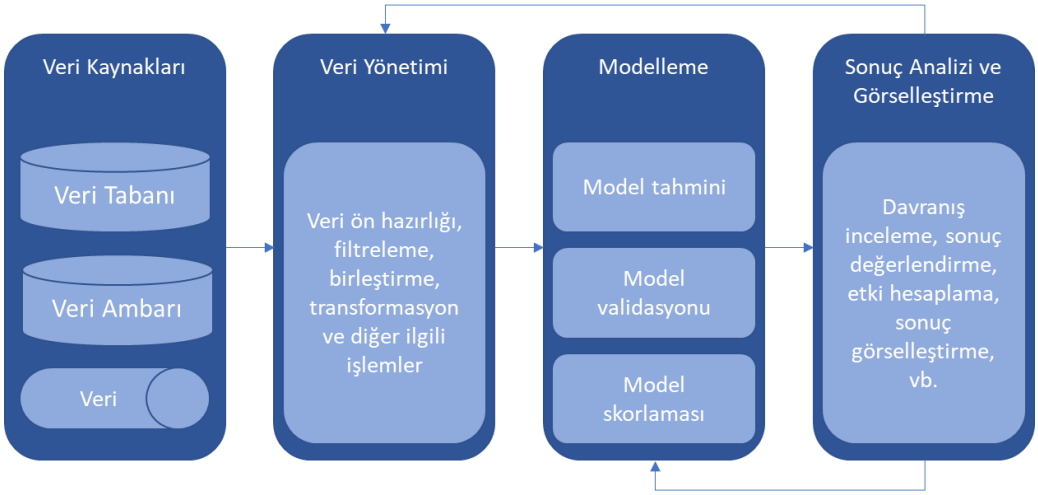
Farklı kaynaklarda veri analitiği projeleri farklı sayılarda adımdan oluşan ama proje süreci benzer projeler olarak karşımıza çıkmaktadır. Gandomi ve Haider (2015) büyük veri analitiği sürecini şu 5 ana bölüm altında gruplandırmışlardır (Gandomi, A & Haider, M., 2015):

- Veri Temini ve Kaydı
- Veri Hazırlama ve Temizleme
- Veri Birleştirme ve Görselleştirme
- Modelleme ve Analiz
- Yorumlama



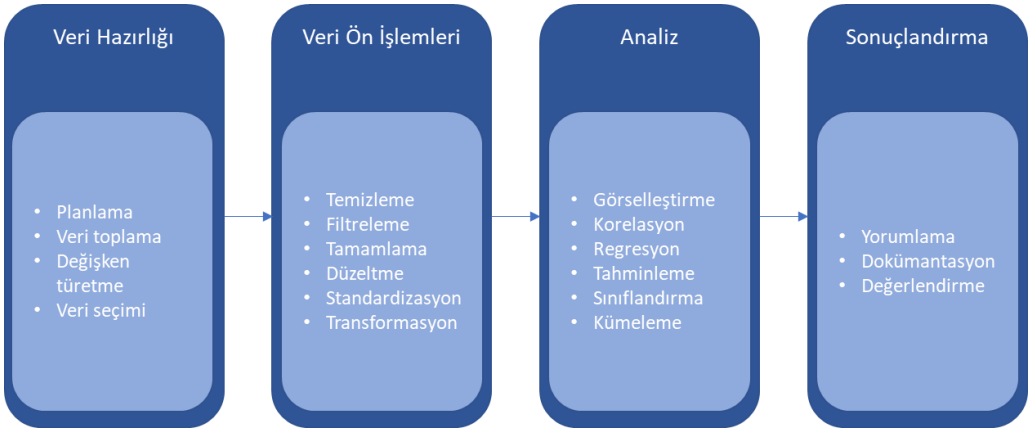
Şekil 4. Veri Analitiği Proje Süreçleri (1)

Assunçao, M., vd. ise veri analitiği süreçlerini 4 ana başlık altında gruplamıştır (2015). Bu gruplama da aslında içerik olarak bir öncekinden çok farklı değildir.



Şekil 5. Veri Analitiği Proje Süreçleri (2)

Runkler ise kitabında bu gruplamayı aşağıdaki şekilde yapmıştır (2016). Bundan sonra da bu gruplama üzerinden detaylandırma yapılacaktır.



Şekil 6. Veri Analitiği Proje Süreçleri (3)

2.3.1. Veri Hazırlığı

Veri hazırlığı adımında yapılan başlıca işlemlerden ilki hangi veri kaynaklarının kullanılacağının, hangi veri dönemlerinin alınacağının, hangi formatta verilerin toplanacağının belirlendiği planlama aşamasıdır. Bu aşama düzgün bir veri analitiği çalışmasının temelini oluştururken burada yapılacak bir hata da projenin ilerleyen safhalarında sıkıntıya yol açabilecek ve tekrar bu adımdan başlamak gerekebilecektir.

Bir sonraki veri hazırlığı adımı verinin toplanmasıdır. Planlaması yapılan veri hazırlığında farklı ortamlardan, farklı zaman aralıklarına ait, farklı formatlarda veriler çekilebilecektir. Verilerin tutulduğu kaynak veri tabanına ve verinin formatına göre veri toplama eforu da kısalıp uzayabilmektedir.

2.3.2. Veri Ön İşlemleri

Modelleme için veri çekimi yapıldıktan sonraki izlenecek ilk adım verinin modellemeye uygun hale getirilmesidir. Bu aşamada yapılacak işlemler proje bazlı değişmekle birlikte şunlar olabilir:

- Tabloların birleştirilmesi
- Tablo yapısının düzenlenmesi
- Tekrarlayan verilerin temizlenmesi
- Boş değerlerin tespit edilmesi ve işlenmesi
- Ekstrem değerlerin tespit edilmesi ve düzeltilmesi
- Hatalı değerlerin tespit edilmesi ve düzeltilmesi
- Kategorik değişkenlerin numerik hale çevrilmesi
- Veri ölçeklendirme

Tabloların birleştirilmesi: Kompleks veri analitiği projelerinde veriler genellikle farklı kişiler tarafından farklı veri tabanlarından farklı yöntemlerle hazırlanmaktadır. Farklı bilgileri içeren bu tabloların birleştirilmesi gerekmektedir ve bu birleştirme de tablolar içerisinde yer alan anahtar alanlar kullanılarak yapılabilir.

Bu aşamada dikkat edilmesi gereken noktalar şunlardır:

- Hazırlanan her tabloda aynı anahtar alanın yer aldığından emin olunmalıdır. Örneğin; bir e-ticaret şirketi müşterilerinin demografik bilgilerinin yer aldığı tablo müşteri seviyesinde iken alışveriş bilgilerinin yer aldığı tablo işlem seviyesinde olabilir. Bu tabloların her birinde müşteri numarası gibi bir bilgi bulunacağı için birleştirme bu alan üzerinden yapılabilir.
- Hazırlanan her tablonun aynı kümeye ait bilgileri içerdiği kontrol edilmeli ve tabloların birleşmesiyle nihai tablonun büyüklüğü kontrol edilmelidir. Örneğin; bir tabloda tüm 2018 yılı verileri varken diğerinde sadece Ocak-Nisan 2018 verileri varsa sadece bu veriler eşleşerek toplam kayıt sayısında azalma olacaktır.

Tabloların yapısının düzenlenmesi: Veri hazırlığındaki en temel konulardan biri de her tablonun erişilebilir ve düzgün olduğunun kontrol edilmesidir.

Bu aşamada dikkat edilmesi gereken noktalar şunlardır:

- Her tablonun erişilebilir olduğu kontrol edilmelidir (Jafari, 2022).
- Tablolar farklı kişiler tarafından hazırlandığı için alan isimlerinin benzer ve anlaşılır olduğundan emin olunmalıdır (Jafari, 2022).
- Modelleme platformlarının isimlendirme kriterlerine dikkat edilerek gerekli düzenlemeler yapılmalıdır. Örneğin; Türkçe karakterler, boşluklar, özel karakterler, tekrarlayan alan isimleri düzenlenmelidir.

Tekrarlayan verilerin temizlenmesi: Modelleme çalışmalarına konu olacak verinin çoklayan ve hatalı şekilde aynı olan kayıtlardan oluşmadığından emin olunmalıdır. Veri içerisinde tesadüfi olarak her bilgisi aynı olan kayıtlar yer alması mümkündür ama bu kayıtlarda dahi kayıt numarası gibi bir anahtar bilginin değişmesi beklenir. Eğer bu bilginin de birebir aynı olduğu görülüyorsa bir hata olması muhtemeldir.

Boş değerlerin tespit edilmesi ve işlenmesi: Veri içerisinde teknik olarak kesintiler sebebiyle o anda veri tutulamadığından ya da o anda ilgili veri kaynağı mevcut olmadığından boş

değerler gelmesi mümkündür. Bunun dışında belli müşteriler için belli bilgilerin dolmaması da söz konusu olabilir. Örneğin; bir banka veri tabanında müşteri bazlı bir veride müşterinin kredi kart bulunmuyorsa kredi kartı ile ilgili bilgiler boş olabilir. Makine öğrenmesi algoritmaları boş değer bulunan satırlar ile işlem yapamamaktadır. Bu sebeple boş değer içeren satırlar için önümüze üç yöntem çıkmaktadır (Brownlee, 2020):

1. Bu yöntemlerden ilki boş değer bulunan satırların tamamen silinmesi olup eğer ancak az sayıda kaydı etkiliyorsa ve kitle içerisinde tutarsızlıklara yol açmayacaksa tercih edilebilir.
2. Bir diğer ise istatistiki eksik veri doldurma yöntemleri ile boş gelen değerlerin doldurulmasıdır. Boş verilerin istatistiki hesaplamalardan ortalama, medyan, mod değerler ile ya da sabit bir değer ile doldurulması en çok tercih edilen yöntemlerdir. Örneğin; bir veri kümesinde müşterilerin bir kısmının yaş bilgisi boş geliyorsa bu müşteriler için medyan yaş değeri ile yaş bilgisi doldurulabilir.
3. Bahsedilen diğer iki yönteme göre daha kompleks bir yöntem olan k-en yakın komşu algoritması ile boş değerler için atanacak değerlere bir model ile karar verilebilir.

Ekstrem değerlerin tespit edilmesi ve düzeltilmesi: Ekstrem değerler, veri içerisinde merkezi eğilimden çok uzakta yer alan değerleri ifade etmektedir. Bunlar medyan değerden uzaklık ile veya ortalamadan 2-3 standart sapma uzaklık ile kısıtlanarak tespit edilmeye çalışılmaktadır (Igual, L., & Seguí, S., 2017).

Ekstrem değerlerin düzeltilmesinde uygulanabilecek bazı yöntemler şunlardır:

- Her alan için ekstrem değer içeren satırların tespit edilmesi ve bu satırların silinmesi bir yöntemdir. Ancak herhangi bir alanında ekstrem değer bulunan çok sayıda alan olabileceğinden bu yöntem çok faz kaydın silinmesi sonucuna yol açabilir ve en ideal yöntem değildir.
- Ekstrem değerlerin düzeltilmesinde daha çok tercih edilen yöntem bu değerlerin ekstrem olmayan en minimum ve maksimum değerler ile değiştirilmesidir. Örneğin; bir alışveriş verisinde fatura tutarının dağılımında minimum değer 0.1 TL, %5'lik dilime gelen değer 50 TL ise 50 TL altındaki değerleri 50 TL şeklinde güncellemek mümkündür. Aynı şekilde maksimum fatura tutarı 300 bin TL ama %99'luk dilimdeki değer 12 bin TL ise bu değer üzerindeki kayıtları bu değer ile değiştirmek mümkündür. Bu düzeltme işleminin her değişken özelinde ayrıca yapılması gerekmektedir.

Hatalı değerlerin tespit edilmesi ve düzeltilmesi: Verideki hatalar rassal hata ve sistematik hata olmak üzere iki grupta incelenebilir. Rassal hatalar beklenmeyen nedenlerle gerçekleşen hatalardır. Örneğin; anket çalışmalarında anlaşmazlıklar ve yanlış sorular sebebiyle doğru olmayan sonuçlar çıkması rassal hata olarak sınıflandırılmaktadır. Diğer taraftan, tüm veri toplama süreci boyunca var olan bir sorun dolayısıyla ortaya çıkan hatalar ise sistematik hata olarak sınıflandırılır. Örneğin; ayarı bozuk bir termometre ile her saat başı oda sıcaklığı ölçüldüğünde düzenli bir hata yapılacaktır ve bu da sistematik bir hatadır (Jafari, 2022). Sistematik hataların önlenmesi için veri toplama sürecinin iyi tasarlanması ve toplanan verinin özet istatistikler yöntemiyle tüm veri toplama süreci bitmeden de incelenmesi önemlidir. Bunun yanında hatalı verilerin düzeltilmesi için mümkünse dış kaynaklardan veri temini veya eldeki verinin veri manipülasyonu yöntemiyle düzenlenmesi de uygulanabilecek yöntemlerdendir.

Kategorik değişkenlerin numerik hale çevrilmesi: Modelleme algoritmaları hem girdi hem de çıktı değişkenlerin numerik olmasını zorunlu kılar. Bu sebeple veri içerisinde yer alan tüm kategorik değişkenler de numerik formata dönüştürülmelidir (Brownlee, 2020).

Kategorik değişkenler üç şekilde karşımıza çıkmaktadır (McNulty, 2021):

- İkili değişken: Evet/hayır, Var/Yok, Doğru/Yanlış gibi iki sınıflı değişkenlerdir.
- Sıralı kategoriler: Düşük/orta/yüksek gibi sıralı ifadeler içeren değişkenlerdir.
- Sırasız kategoriler: Araba/Bisiklet/Tren gibi bir sıralama ifade etmeyen değişkenlerdir.

Buna göre kategorik değişkenlerin numerik formata getirilmesi için şu iki yöntem uygulanır:

- One-hot encoding: Bu yöntem ile her bir kategorik değişken ikili değişkenlere dönüştürülür (Brownlee, 2020). Örneğin; cinsiyet değişkeni için her bir kayıtta erkek veya kadın şeklinde iki seçenek olduğunu düşündüğümüzde bu değişkene one-hot encoding uygulandığında 0-1 değerlerini alarak erkek olma durumunu gösteren bir değişken ve 0-1 değerlerini alarak kadın olma durumunu gösteren bir değişken üretilmiş olur. Eğer kategorik bir değişken içerisinde onlarca, yüzlerce farklı değer yer alıyorsa bu değişkene one-hot encoding dönüşümünü uygulamak çok fazla sayıda ikili değişken üretilmesine yol açacağından veri boyutunu çok yükseltebilir.
- Label encoding: Bu yöntemde ise kategorik değişkende yer alan her tekil değere bir sayı karşılığı verilir. Örneğin; bir veri kümesinde meslek değişkenleri yer alıyor ve 50 farklı meslek bilgisi varsa her bir mesleğe 0,1,2, …, 49 şeklinde bir sayı karşılığı atanır.

Veri ölçeklendirme: Birçok makine öğrenmesi algoritması ölçeklendirilmiş veri ile daha iyi performans göstermektedir. Bu ölçeklendirme yöntemleri normalizasyon ve standardizasyon olarak ikiye ayrılır. Normalizasyon tekniğinde, her girdi değişken 0-1 arasında yer alacak şekilde düzenlenir (Brownlee, 2020).

$$y = \frac{x - \text{minimum}}{\text{maksimum} - \text{minimum}}$$

Standardizasyon tekniğinde ise her değerden ortalama değer çıkarılarak standart sapmaya bölme işlemi yapılarak ortalaması 0 ve standart sapması 1 olan bir dağılım elde edilir (Brownlee, 2020).

$$y = \frac{x - \text{ortalama}}{\text{standart sapma}}$$

Analiz

Modelleme için veri çekimi yapılıp veri ile ilgili ön düzenlemelerin tamamlanmasının ardından yapılması gereken işlemler şunlardır:

- Veri görselleştirme
- Veri mühendisliği
- Korelasyon kontrolleri
- Veri veya boyut azaltma
- Modelleme

Veri görselleştirme: Veriyi sadece ortalama, medyan, mod gibi istatistiklerine bakarak anlamak zordur ve tehlikeli olabilir. Çok farklı veriler aynı istatistiklere sahip olmalarına rağmen dağılımları çok farklı olabilir. Bu sebeple dağılımların da incelenmesi gerekli bir adımdır (Igual, L., & Seguí, S., 2017).

Modelleme öncesinde bir veri kümesinde bakılması gereken görselleştirmeler aşağıdaki şekildedir (Jafari, 2022):

- Popülasyonun anlaşılması: Veri içerisindeki her bir değişken için numerik değişkenlerde histogram, çubuk grafik, kutu grafiği gibi grafikler ile kategorik değişkenlerde ise dilim grafik veya histogram grafikler ile popülasyonun özellikleri ortaya çıkarılabilir.
- Popülasyonların karşılaştırılması: İki farklı popülasyon yine benzer grafikler ile karşılaştırılabilir. Bu popülasyonlar model geliştirme, test ve validasyon gibi aynı

modellemeye konu olan farklı alt popülasyonlar da olabilir, modelleme kitlesine dahil olmayan farklı bir kitle ile karşılaştırma da söz konusu olabilir.

- İki girdi değişken veya girdi değişkenler ile hedef değişken arasındaki ilişkinin tespiti: Bu uygulama ile değişkenlerin birbiri ile ilişki düzeyi görsel olarak sunulabilir. Görselleştirme için çok sayıda grafik tipi olmakla birlikte kullanılacak grafiğin hangisi olacağı değişkenlerin formatına bağlıdır. İki numerik değişken arasındaki ilişki serpme diyagramı ile, iki kategorik ikili değişken arasındaki ilişki ısı haritası ile, iki kategorik ancak ikili olmayan değişken arasındaki ilişki kontenjan tablosu ile, bir numerik ve bir kategorik değişken arasındaki ilişki ise kontenjan tablosu veya serpme diyagramı ile görselleştirilebilir.
- Görsel boyutlar ekleme: Şu ana kadar bahsedilen görsel grafiklerin hepsi iki boyutludur ancak veri analistinin veri içerisindeki örüntüleri anlaması için daha fazla boyutu incelemesi gerekecektir. Bu sebeple de iki boyutlu bir serpme diyagramı farklı kategoriler için çoklayarak, filtreler ekleyerek veya renk ve büyüklük boyutu ekleyerek yeni boyutlar da dikkate alınabilmektedir.
- Trendlerin ortaya çıkarılması ve karşılaştırılması: Özellikle zaman serisi analizlerinde olmakla birlikte zaman içerisindeki örüntüleri anlamak amacıyla da grafikler kullanılmaktadır. Örneğin; bir e-ticaret sitesinin satışlarının zamanla değişimi için bir çizgi grafik kullanılabilir veya bir bankanın batık kredilerinin ay ay değişimini görmek için bir çubuk grafik kullanılabilir.

Veri mühendisliği: Her zaman ham verideki değişkenler modelleme ve analiz yapmaya yeterli olamamaktadır. Bu sebeple, ham veri üzerinden yeni ve anlamlı olabilecek değişkenlerin türetilmesi faydalıdır ve bazen de zorunludur. Örneğin, günlük veriden müşteri seviyesinde bir veri hazırlamak veya müşteri seviyesindeki veri içerisinde değişkenleri çaprazlayarak yeni değişkenler türetmek bu aşama içerisinde değerlendirilmektedir. Bu yöntem ile ham veride yer alan değişkenler dışında birçok yeni değişken de modellemede dikkate alınabilmekte ve anlamlı çıkabilmektedir. Son yıllarda otomatik veri mühendisliği yöntemleri ile bu iş de otomasyona tabi tutulabilmiştir.

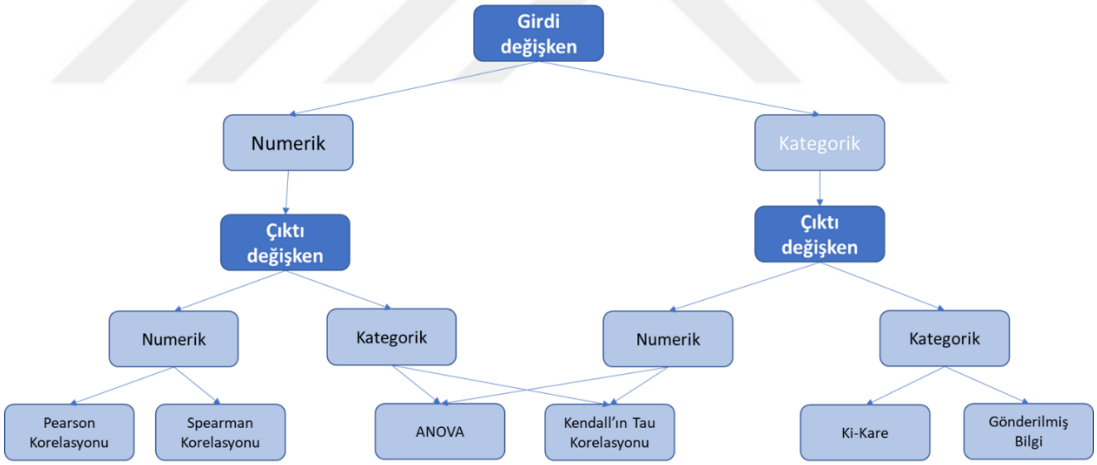
Korelasyon kontrolleri: Veriyi görselleştirme dışında veri içerisindeki alanları istatistiki testler ile de değerlendirmek mümkündür. Bu kontrollerden en önemlisi korelasyondur ve hem girdi değişkenler arasında korelasyon hem de hedef değişken ile her bir girdi değişken arasındaki korelasyonu incelemek önemlidir.

Kovaryans, ölçümlenen iki değişkenden birisi değiştikçe diğerinin ne kadar değiştiği bilgisini sunar. Korelasyon ise kovaryansın iki değişkenin standart sapmalarının çarpımına bölünmesidir. 0-1 arası değer alan korelasyon 1 değerine ne kadar yakın ise bahsi geçen iki değişken arasındaki lineer ilişki o derecede yüksektir. Pearson'ın rho korelasyonu ve Kendall'ın thu korelasyonları en sık kullanılan iki korelasyon tekniğidir (McNulty, 2021).

Veri/boyut azaltma: Bu adımın uygulanma sebepleri şunlardan biri veya birkaçı olabilir:

- Modelleme için kullanılan veri sayısının çok yüksek sayıda olması ve buna bağlı olarak modelin çalışma süresinin yüksek olması, bunun teknik olarak mümkün olmayan bir zaman olması veya ticari kısıtların üzerinde bir süre olması
- Gelişmiş makine öğrenmesi algoritmalarında iki değişken arasındaki ilişki çok baskın şekilde nihai sonucu etkilemese de özellikle geleneksel modelleme algoritmalarında lineer ilişkideki baskının boyutunu azaltma ihtiyacı olması

Veri boyutu azaltma teknikleri ve nihai değişken seçiminde kullanılan teknikler şu şekilde listelenebilir (Brownlee, 2020):



Şekil 7. Veri Boyutu Azaltma Teknikleri

İstatistiksel nedenler dışında farklı sebeplerle de modelleme verilerinde eleme yapıldığı da sektörde gözlemlenmektedir. Bunların bazıları şu şekilde sıralanabilir:

- Girdi değişkenlerden beyana dayalı olan ve doğruluğundan emin olunamayan değişkenler. Örneğin; bankacılık sektöründe dijital kanallardan yapılan kredi başvurularında alınan meslek veya maaş bilgileri bu sınıfta yer almaktadır.

- Geçmişte bir süre doğru beslenmediği bilinen veya doğruluğundan şüphe duyulan veriler. Örneğin; telekomünikasyon sektöründe anlık olarak toplanan baz istasyonu verilerinden elde edilen konum bilgileri.
- Gelecekte çeşitli sebepler ile toplanmaması düşünüldüğünden modelde yer almaması istenen veriler. Örneğin; dijital kanallarda müşteri deneyimi sebebiyle daha az soru sormak ve daha az bilgi toplamak şeklinde bir trend olduğundan müşteriden alınan kısıtlı bilgi ile kredi teklifi çıkmayı planlayan bir dijital bankanın artık meslek bilgisini toplamayacağı için bu bilgiyi elemesi.
- Adalet ilkesine aykırı olduğu gerekçesi ile bazı değişkenlerin elenmesi. Özellikle Avrupa'da cinsiyet, lokasyon, milliyet gibi bilgilerin modellerde manuel şekilde elenmesi.

Modelleme: Modelleme için gerekli veri hazırlanıp ön işlemler yapıldıktan ve analizler tamamlandıktan sonra yapılması gereken işlem ilgili probleme uygun modelleme algoritmaları kullanılarak modelin geliştirilmesidir. Modelleme algoritmalarına ayrı bir başlık altında daha detaylı değinileceği için bu bölüm kısa tutulmuştur.

2.3.3. Sonuçlandırma

Modelleme çalışmasının tamamlanmasının ardından izlenmesi gereken adımlar şunlardır:

- Bağımsız Validasyon
- Simülasyon ve sonuç değerlendirme
- Modelin canlıya alınması
- Dokümantasyon

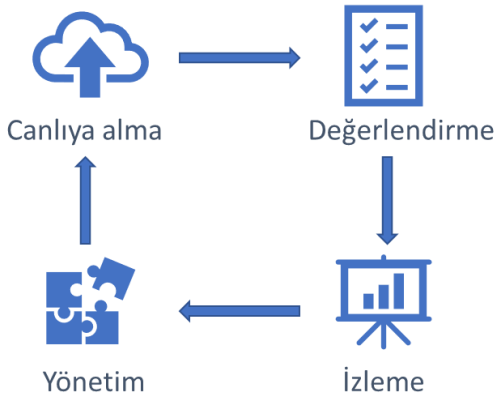
Bağımsız Validasyon: Modelleme sırasında modelleme kitlesinin geliştirme, test ve validasyon dışında ayrılması şeklinde bir validasyon yapılmaktadır. Ancak bu yöntem sadece model performansı anlamında bir kontrol yetkinliği getirmektedir. Bunun dışında ayrıca bağımsız bir validasyon ekibi veya personelinin izlenen proje yöntemini tüm adımları ile tekrar etmesi, doğrulaması ve tüm sürecin tekrarlanabilir olduğunu kontrol etmesi gerekmektedir. Özellikle bankalarda geliştirilen modeller için bu süreci yürütecek ayrı ekipler görevlendirilmektedir ve model geliştiren ekiplerin modellerini ayrı bir ekip kontrol edip onay vermektedir.

Simülasyon ve sonuç değerlendirme: Geliştirilen modelin yapılan iş üzerindeki ana performans metriklerine etkisinin ölçümlenmesi adımıdır. Bu adımda modelin teknik performansı dışında değerlendirilmesi gereken konular değerlendirilir. Örneğin; bir dolandırıcılık ve suiistimal tespit

modeli geliştiriliyorsa bu modelin performansı sadece Gini katsayısı ile değil aynı zamanda "false positive" kayıt sayısındaki azalma ile de ölçülür. Dolayısıyla bunun gibi metriklerin geçmiş veride yapılan simülasyonlar ile hesaplanması önemlidir.

Modelin canlıya alınması: Modelin implemente edilmesi ve kullanıma geçmesi de modelleme ardından yapılan analizler sonucunda yapılır. Eğer modelin kullanımı ile ilgili herhangi bir çekince görülmüyorsa bu işlem başlatılmalıdır. Modeller canlı sistemlerde iki türlü çalışabilir. Bunlardan biri çevrimiçi ve anlık şekilde olan yöntemdir ve her model çalışması tekil kayıt için gerçekleştirilir. Örneğin; bankalar müşterilerinin bireysel kredi başvurusu sırasında aynı müşteri için kredi kartı eğilim modellerini de çalıştırarak müşteriye ek ürün sunup sunmama kararını verirler. Bu modeller her müşteri için başvuru sırasında tek tek çalışmaktadır. Diğer bir model çalıştırma yöntemi ise toplu liste halinde tüm müşteriler için modeli bir seferde çalıştırmadır ve bu yöntemde zamanla yarışılmadığından modelin çalışma performansı daha az kritiktir. Örneğin; bir Telekom operatörünün her hafta sonu mevcut müşterilerinin telefon hattını kapatma ihtimalini hesaplaması bu yöntemle olur. Modeller hafta sonu veya ay sonu gibi belirli dönemlerde tüm mevcut müşteriler için çalışacağından saliseler içerisinde cevap verilip aksiyon alınması şart değildir.

Modeller Web servis şeklinde veya Docker veya Kubernetes sistemleri üzerinden canlıya alınabilirler. Modelin canlıya alınması sonrasında düzenli olarak izlenmesi de kritiktir ve canlıya alım sonrasındaki süreç aşağıdaki şekilde çizilebilir (Singh, 2021):



Şekil 8. Modellerin Canlıya Alınma Süreci

Dokümantasyon: Tamamlanan veri analitiği projesinin uçtan uca dokümante edilmesi ve ilgili kişilere aktarımının yapılması iş sürekliliği açısından önemlidir. Bir proje dokümanında özellikle şu başlıkların yer alması önem arz etmektedir:

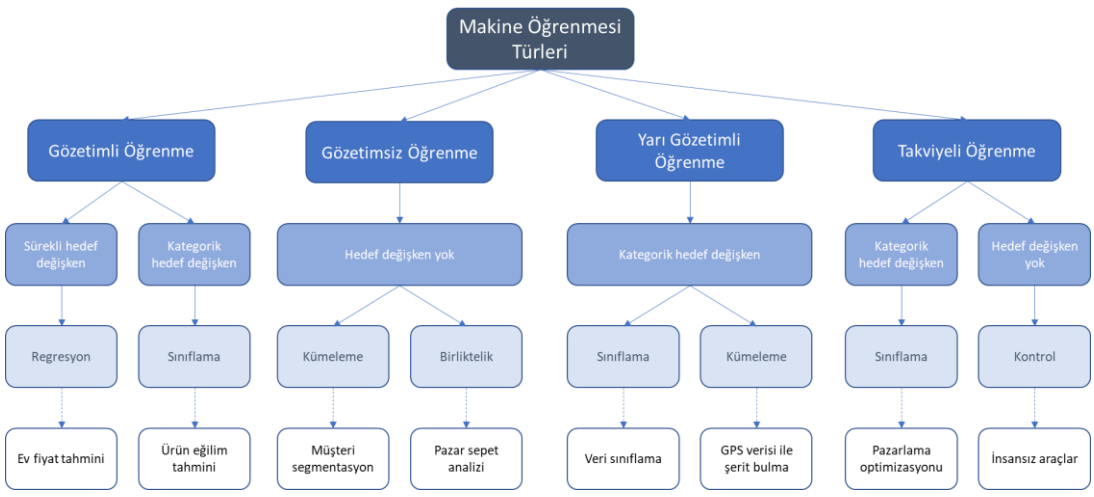
- Proje amacı

- Proje ekibi
- Veri hazırlığı (Kaynaklar, veri hazırlama adımları, yapılan varsayımlar, veri düzenlemeleri, vb.)
- Modelleme detayları (Denenmiş modelleme algoritmalarının detayları, kullanılan parametreler, model sonuçları, vb.)
- Sonuçlar (Model performans sonuçları ve iş performans metrikleri)
- Proje onayı

2.4. Modelleme Algoritmaları

Makine öğrenmesi çözülmeye çalışılan problem tipine göre 4 ana kategoride incelenebilir ve her biri için hedef değişkenin varlığı ve türüne göre uygulanacak modelleme algoritmaları farklılık gösterir (Singh, 2021):

- Gözetimli öğrenme
- Gözetimsiz öğrenme
- Yarı gözetimli öğrenme
- Takviyeli öğrenme



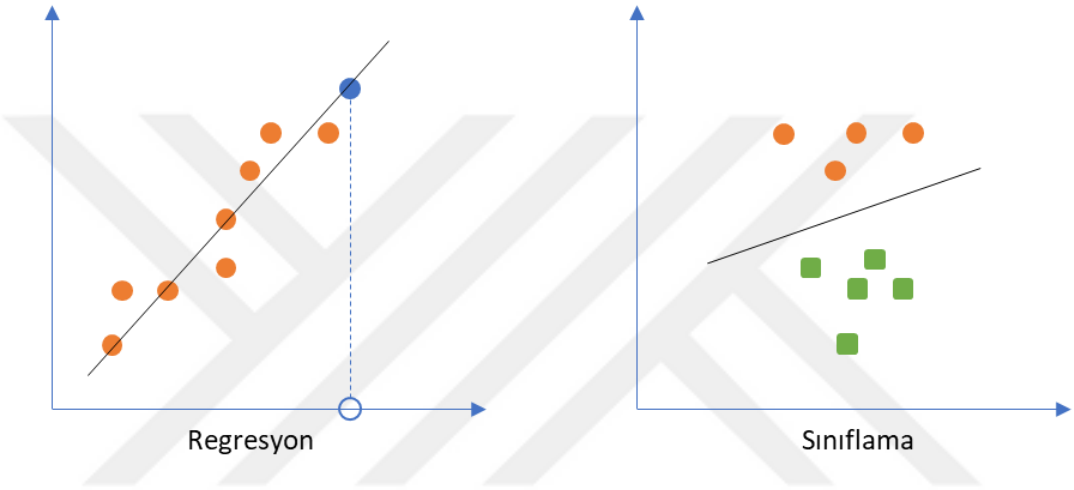
Şekil 9. Makine Öğrenmesi Türleri

2.4.1. Gözetimli Makine Öğrenmesi

Bu makine öğrenmesi kategorisi en sık uygulanan makine öğrenmesi kategorisidir ve birçok gerçek hayat uygulaması mevcuttur. Özetle, hedef değişkenin bilindiği ve girdi değişkenler ile bu hedef değişkenin arasındaki ilişkinin ortaya çıkarılmaya çalışıldığı öğrenme şeklidir. Bu tür modellerin

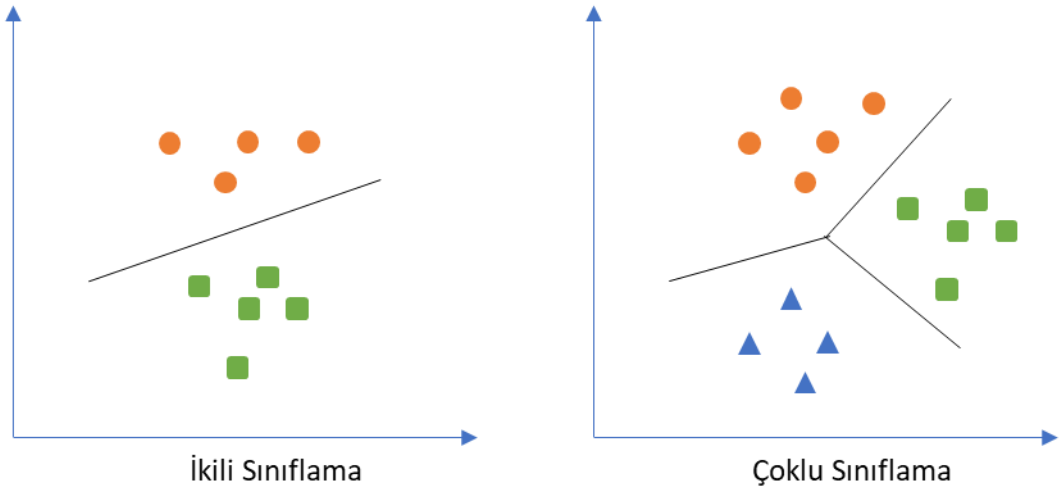
performans göstergesi çok basit ifadeyle gerçek hedef değişken ile tahmin edilen değer arasındaki farka dayanmaktadır ve gerçekleşen ve tahmin edilen değerler arasındaki hatayı optimize edecek fonksiyon bulunmaya çalışılmaktadır. Daha önce görülmüş ve çıktısı bilinen kayıtlardan hareketle oluşturulan bu fonksiyon kullanılarak daha önce görülmemiş kayıtların çıktısı tahmin edilmektedir (Singh, 2021).

Gözetimli öğrenme için tarihsel veri mevcuttur ve hedef değişkenin türüne göre iki farklı alt sınıfta değerlendirilmektedir. Eğer hedef değişken sürekli ve numerik bir değişken ise regresyon, hedef değişken kategorik bir değişken ise sınıflama gözetimli öğrenmesi olarak isimlendirilir (Singh, 2021).



Şekil 10. Regresyon ve Sınıflama Problemleri

Sınıflama problemleri de hedef değişkenin ikili veya çoklu sınıftan oluşmasına göre yine alt gruplara ayrılmaktadır. Örneğin; evet/hayır veya var/yok gibi bir sınıflandırma söz konusu ise ikili sınıflama, evet/hayır/belki gibi bir sınıflama söz konusu ise çoklu sınıflama probleminden bahsedilmektedir.



Şekil 11. İkili ve Çoklu Sınıflama Problemleri

Gözetimli öğrenme için kullanılan başlıca makine öğrenmesi algoritmaları şöyle sıralanabilir:

- Lineer regresyon
- Yapay sinir ağları
- Lojistik regresyon
- K-en yakın komşu
- Karar ağacı
- Naive Bayes sınıflandırıcısı
- Lineer diskriminant analizi
- Destek vektör makineleri
- Vektör nicemleme
- Ekstrem gradyan artırma
- Rassal ağaç

Bu algoritmalardan en önemli olan ve sektörde kullanımı yaygın olanlarından birkaçını detaylı inceleyebiliriz.

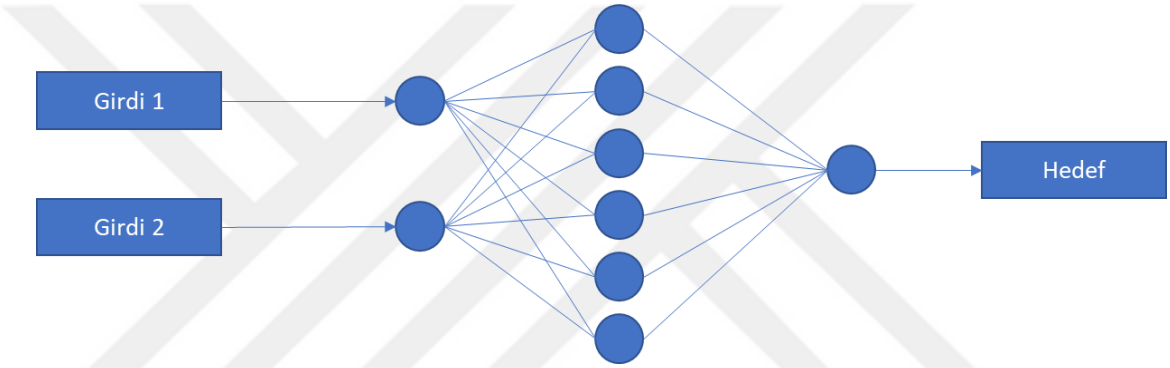
Lineer Regresyon: Lineer regresyon, değişkenler arasında bir ilişki bulabilmek için genel bir eşitlik kullanır ve bu eşitlik ile hedef değişkeni (bağımlı değişken) bağımsız değişkenlerden tahmin edecek modeli belirler. Lineer regresyonun genel eşitliği şu şekildedir:

$$target = \beta_0 + \beta_1 \, x \, predictor1 + \beta_2 \, x \, predictor2 + \cdots + \beta_N \, x \, predictorN$$

Bu eşitlikte N bağımsız değişken sayısını vermektedir. Lineer regresyonun çalışma metodu basittir ve öncelikle eşitliği en iyi şekilde sağlayan β değerleri tahmin edilir ve bu eşitlik de görülmemiş kayıtların tahminlenmesinde ve skorlanmasında kullanılır (Jafari, 2022).

Lineer regresyonun sınıflama problemlerinin çözümünde kullanılması da mümkün olmakla birlikte çok iyi sonuç vermezler. Bunun sebebi de hataların karesinin minimize edilmesi şeklindeki yaklaşımın sınıflama problemlerine çok uygun olmamasıdır (Jiang, 2021).

Yapay Sinir Ağları: Lineer regresyon problem çözümünde bir eşitlik kullanırken yapay sinir ağlarında ise bağımsız değişkenler ile bağımlı değişkeni birbiri ile ilişkilendirmek için nöronlardan oluşan bir ağ kullanılır. Bu ağın bir örneği aşağıda sunulmuştur (Jafari, 2022).



Şekil 12. Yapay Sinir Ağları

Bu görselde görülen yuvarlak şekillerin her biri nöronlardır. Bu nöronlar girdi katmanında, çıktı katmanında veya gizli katmanda olabilir. Girdi katmanında her bir bağımsız değişken için bir nöron yer alır ve bu örnekte 2 adet bulunmaktadır. Çıktı katmanındaki nöronlar ise genellikle bir tane olur ve hedef değişkeni temsil eder. Bu örneğimizde de 1 adet çıktı nöron bulunmaktadır. Girdi ve çıktı katmanlar arasında ise gizli katman yer alır ve burada da nöronlar bulunur. Gizli katmanların büyüklüğü ve bu katmandaki nöronların sayısı modelin karmaşıklığına ve hesaplama maliyetine göre belirlenmekte ve çeşitlenebilmektedir. Nöronların her birini birbirine bağlayan ve görselde oklar ile görünen her bir ilişkiye de bağlantı adı verilir (Jafari, 2022).

Yapay sinir ağlarının lineer regresyona benzer yönleri olduğu kadar farklı yönleri de vardır. Öncelikle ortak noktalarını sıralayabiliriz (Jafari, 2022):

- Lineer regresyon yapısal bir eşitlik ile bağımsız ve bağımlı değişkenler arasındaki ilişkiyi ifade etmeye çalışır. Aynı şekilde yapay sinir ağları da ağ yapısı ile benzer ilişkiyi yakalamaya çalışır.

- Lineer regresyonda β ağırlıkları belirlenerek en iyi eşitlik elde edilmeye çalışılır. Benzer şekilde yapay sinir ağlarında da bağlantıların her birinin ağırlığı bir parametre olarak değerlendirilerek en iyi bağlantılar bulunmaya çalışılır. Her iki yöntem de veriyi kullanarak kendilerini günceller ve en iyi sonucu yakalamak hedeflenir.
- Her iki algoritmada da değişken katsayıları veya bağlantıların ağırlıkları belirlenerek model elde edildikten sonra bu algoritma yeni kayıtların değerlendirilmesinde kullanılır.

Yapay sinir ağlarının lineer regresyona göre farklılaştığı noktalar da vardır (Jafari, 2022):

- Lineer regresyonda değişkenlerin ağırlıkları belirlendikten sonra karşımıza basit denilebilecek bir eşitlik çıkmaktadır. Öte yandan, yapay sinir ağlarında daha fazla gizli katman veya nöron kullanılarak model karmaşık hale gelebilir. Böyle bir modelin daha karmaşık problemlerin çözümünde etkisi olur ancak hesaplama maliyetini de artırır.
- Lineer regresyon formülasyonu tamamen matematiksel bir altyapıya dayanırken sinir ağlarında sezgisel bir taban da bulunmaktadır.

Lojistik Regresyon: Lojistik regresyon algoritması olasılıksal istatistiksel sınıflama problemleri için kullanılır. Bağımlı değişkenin ikili bir değişken olduğu durumlarda birden fazla bağımsız değişkeni dikkate alarak çalışır. Lojistik fonksiyonun genel şekli şu şekildedir:

$$f(x) = \frac{1}{1 + e^{-kx}}$$

Lojistik fonksiyon formülü girdi negatif veya pozitif şekilde sonsuza gitse bile 0-1 arasında sonuç üretmeyi mümkün kılar. Lojistik regresyon algoritmasının uygulandığı başlıca problemler bankacılıktaki temerrüt modelleri, ürün eğilim modelleri, müşteri terk etme modelleri gibi sayılabilir (Igual, L., & Seguí, S., 2017).

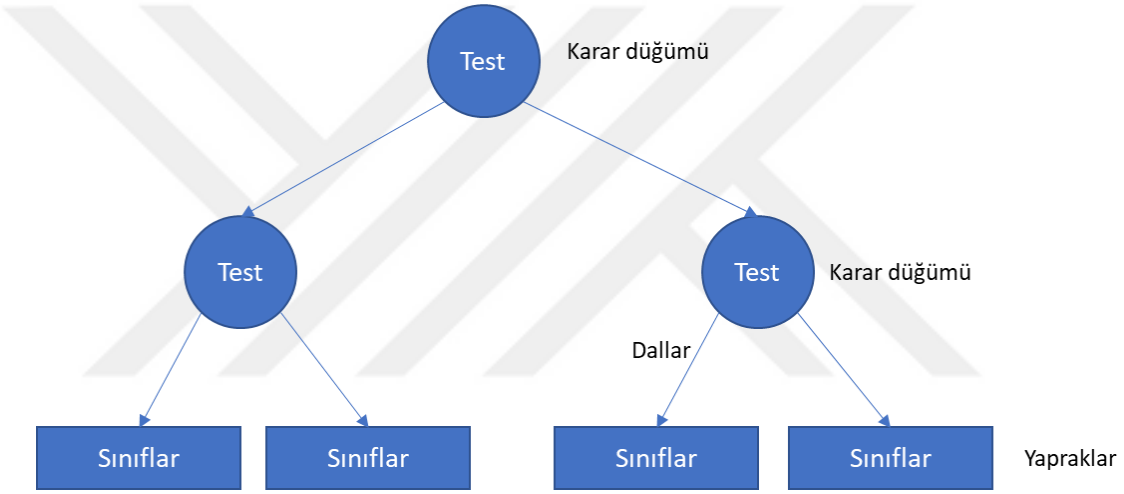
K-en yakın komşu: K-en yakın komşu algoritması da makine öğrenmesi algoritmalarının en önemli üyelerinden biridir. Hem numerik bir hedef değişkenin olduğu regresyon problemlerinde hem de kategorik bir hedef değişkenin olduğu sınıflama problemlerinde kullanılabilir. Algoritmanın temelinde her bir değişken arasındaki en kısa mesafeyi bulmak ve bunları birleştirdikten sonra başka değişkenler ile yine en kısa mesafeleri hesaplaya hesaplaya ilerlemek vardır. Bu mesafe hesaplaması Öklid mesafesidir ve formülü aşağıdaki şekildedir (Jung, 2022).

$$\text{Öklid mesafesi} = \|x - x'\|$$

En yakın komşu algoritmalarında bir sınıflama fonksiyonu oluşturulmaz ama model geliştirme verisi kaydedilir ve sınıflama isteği geldiğinde değerlendirilir. Buna tembel öğrenme denir. Bu sebeple en yakın öğrenme algoritmalarının en önemli avantajı yeni gelen bir verinin anında dahil edilebilmesi ve lineer olmayan sınırların da tanımlanabilmesidir. Öte yandan en önemli dezavantajı ise karmaşık hesaplama yapısıdır (Runkler, 2016).

Karar ağacı: Karar ağaçları ağaç şeklinde yapılardır ve sınıflama ve regresyon problemleri için kullanılabilirler. Karar ağaçları akış şeklindeki çizimleri ile kuralların görselleştirilmiş halidir ve bu yöntem ile geliştirilen modeller en açıklanabilir modeller arasındadır (Alao, Adeyemo, 2013).

Karar ağacının görselleştirilmesi aşağıdaki şekilde yapılabilir (Kozak, 2019):



Şekil 13. Karar Ağacı

Karar ağaçlarında dikkate alınması gereken iki düğüm vardır (Kozak, 2019 & Jung, 2022):

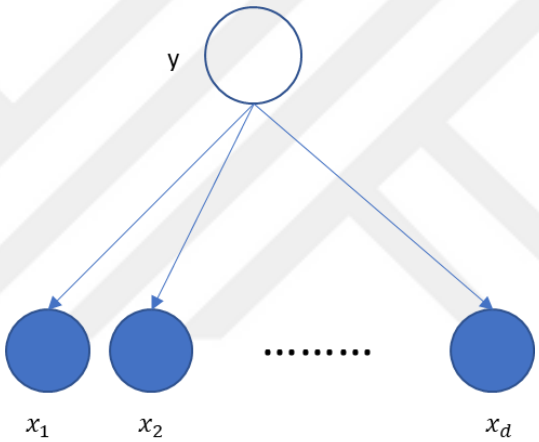
- karar düğümü: değişken vektör ile ilgili testleri ifade eden düğümdür
- yaprak düğümü: değişken uzayının alt kümelerini ifade eden düğümdür

Karar ağaçları anlaşılması en kolay algoritmalardan olmasına rağmen modelin geliştirilmesinde dikkat edilmesi gereken belli parametreler ve noktalar vardır:

- Her bir düğüme düşen minimum kayıt sayısı kontrol altında tutulmalıdır ve çok az sayıda kaydın bir düğümde yer alması ile aşırı uyma problemi yaşanmamalıdır.
- Ağacın aşağı doğru kaç adım gideceğine de dikkat edilmelidir. Adım sayısı arttıkça karmaşıklık artmakta ve en alttaki düğümlere düşen kayıt sayısı azalmaktadır. Bu da yine aşırı uyma problemine yol açabilir.

Karar ağaçlarında ağacın dallara bölünmesi sırasında kullanılan algoritmalar CART ve C4.5 algoritmalarıdır. Bu algoritmalar ile boş değerler, sürekli değişkenler, büyük veri kümeleri gibi zorluklar da çözülebilmektedir ve CART algoritması ile her bir bölünmede Gini gibi bir performans metriği kontrol edilerek adım adım ilerlenmektedir. Bu sebeple karar ağaçlarının oluşturulması "böl ve yönet" şeklinde de örneklenmektedir.

Naive Bayes sınıflandırıcısı: Sınıflandırma problemlerinde amaç bilinmeyen kaydı bilinen k sınıftan birine $y \in \{w_1, w_2, \ldots, w_k\}$ atamaktır. Bu yapılırken de d adet değişken $\{x_1, x_2, \ldots, x_d\}$, girdi olarak kullanılmaktadır. Naive Bayes koşullu olasılık teoremine göre bu değişkenler, y sınıfı için birbirinden bağımsızdır. Naive Bayes teoreminin en basit görünümü olan Bayesian ağı aşağıdaki şekilde görülebilir (Jiang, 2021):



Şekil 14. Naive Bayes Sınıflandırıcısı

Naive Bayes teoremine göre y sınıfı d adet değişkene bağımlı olup teorem aşağıdaki şekilde formüle edilmektedir (Jiang, 2021):

$$p(y, x_1, x_2, \ldots, x_d) = p(y)p(x_1|y)p(x_2|y) \ldots . p(x_d|y)$$

$$= p(y) \prod_{i=1}^{d} p(x_i|y)$$

Naive Bayes sınıflandırıcı algoritmasında ise, tüm bilinen değerler bilinmeyen y sınıfını belirlemek için kullanıldığından sınıflandırıcı algoritma şu şekilde formüle edilmektedir (Runkler, 2016):

$$y^* = \arg \max_y p(y| x_1, x_2, \ldots, x_d) = \arg \max_y p(y) \prod_{i=1}^{d} p(x_i|y)$$

Naive Bayes sınıflandırıcı esnek bir yapıya sahiptir ve değişkenin tipine göre farklı dağılımlar kullanılabilir. Örneğin; x_i ikili bir değişken ise Bernoulli dağılımı, ikili olmayan bir kesikli değişken ise multinomial dağılım, sürekli bir değişken ise Gaussian dağılımı kullanılabilir (Jiang, 2021).

Lineer diskriminant analizi: Lineer diskriminant analizi (LDA), çoklu sınıf sınıflandırma problemleri için kullanılan bir makine öğrenmesi algoritmasıdır. LDA, geliştirme veri kümesini en iyi şekilde ayırmaya çalışmak için kullanılan bir yöntemdir. Girdi değişkenlerin lineer kombinasyonları kullanılarak sınıflar arasında maksimum fark ve her sınıf içerisindeki kitle içerisinde de (merkezi nokta veya ortalama) minimum fark yakalanmaya çalışılmaktadır. Literatürde yapılan bir başka tanımda ise grup içerisindeki varyansı minimize ederken, gruplar arasındaki varyansı maksimize eden yöntem olarak tanımlanmaktadır (Brownlee, 2020).

Sınıflar arasındaki maksimum farkı bulmak için formülasyon aşağıdaki şekilde verilmektedir (Jiang, 2021). Tüm yüksek boyutlu vektörlerin K sınıftan geldiğini varsayalım. Öncelikle vektörleri K alt kümeye ayıralım $\{C_1,\ C_2, \ldots,\ C_K\}$. Sonra, ortalama vektörü ve kovaryans matrisini hesaplayalım.

$$\mu_k \quad = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i$$

$$S_k \quad = \frac{1}{|C_k|} \sum_{x_i \in C_k} (x_i - \mu_k)(x_i - \mu_k)^T$$

Tüm vektörler lineer bir doğrultudaki bir w vektörü formunda ifade edilebilir. Sınıfların birbirinden farklarının maksimum olmasından bahsedildiği noktada aynı sınıftaki tüm projeksiyonlar birbirine çok yakın durumda olmalıdır ve farklı sınıflar da uzak noktalarda olmalıdır. Bu iki amacın aynı anda gerçekleşebilmesi için Fisher'ın lineer diskriminant analizi aşağıdaki w projeksiyonunu maksimiz etmeyi hedefler.

$$J_w = \max_w \frac{w^T S_b\, w}{w^T S_w\, w}$$

Bu oranda pay kısmı, sınıflar arası farkları ifade etmektedir ve sınıflar arası serpilme matrisi S_b kullanılarak hesaplanmaktadır. Burada S_b tüm ortalama vektörler üzerinden hesaplanmaktadır ve μ tüm vektörlerin ortalama vektörüdür.

$$S_b = \sum_{k=1}^{K} |C_k|\ (\mu_k - \mu)(\mu_k - \mu)^T$$

Paydada ise sınıflar içi yakınlık hesaplanmaya çalışılmaktadır ve sınıf içi serpilme matrisi S_w kullanılmaktadır. S_w tekil örneklem kovaryans matrislerinin toplamıdır.

$$S_w = \sum_{k=1}^{K} S_k$$

Bu bilgiler ışığında, J_w matrisini maksimize etmek şeklindeki optimizasyon problemi şu şekilde ifade edilebilir:

$$w^* = \arg \max_w w^T S_b \, w$$

$$kısıt: \qquad w^T S_w \, w = 1$$

Destek vektör makineleri: Destek vektör makineleri ikili sınıfa sahip sınıflandırma problemleri için kullanılan ve y hedef değişkenini bulmak için x girdi değişkenlerini kullanan bir makine öğrenmesi algoritmasıdır. Destek vektör makineleri de lineer ve lojistik regresyonda olduğu gibi lineer hipotez uzayını baz alır ve karar destek makinelerinin tek farkı kayıp fonksiyonudur (Jung, 2022).

Aşağıdaki görselde D veri kümesi, $x_i \in R^2$ değişken vektörlerinden oluşuyor olsun ve her bir nokta koordinat düzleminde bir noktayı temsil etsin. D veri kümesi iki sınıfa ayrılabilir:

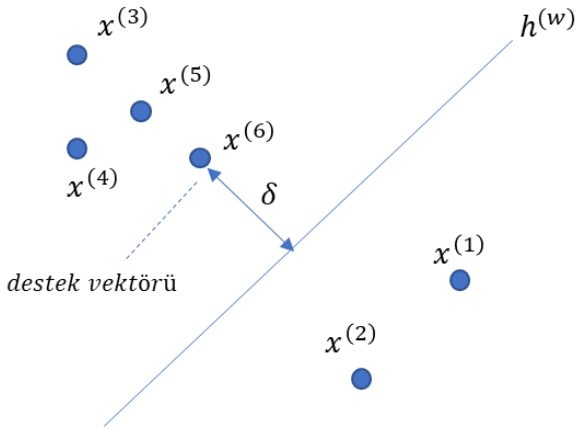
$$C^{(y-1)} = \{x^{(i)} : \, y^{(i)} = 1\}, ve \; C^{(y--1)} = \{x^{(i)} : \, y^{(i)} = -1\}$$

Destek vektör makinesi, bu ayrımı en iyi şekilde yapacak lineer haritayı $h^{(w)}(x) = w^T x$ bulmaya çalışır.

$$h(x^{(i)}) > 0, \qquad x^{(i)} \in \; C^{(y=1)} \; için$$

$$h(x^{(i)}) < 0, \qquad x^{(i)} \in \; C^{(y=-1)} \; için$$

Destek vektör makineleri minimum ortalama hata ile $h^{(w)}$ hipotezini öğrenmeye çalışır. Hatanın minimize edilmesi ise $h^{(w)}$ karar sınırı ile her geliştirme verisi sınıfı arasındaki marj olan δ ‘nın maksimize edilmesidir (Jung, 2022).



Şekil 15. Destek Vektör Makineleri

Destek vektör makinelerinin en basit hali hiperdüzlemi lineer şekilde ayırarak maksimum marja ulaştığı versiyonudur. Bunun daha karmaşık versiyonlarından ilki ise yumuşak karar destek makineleridir (soft SVM) ve lineer olarak ayrılamayan düzlemler için kullanılır. Düzlem lineer olarak ayrılamıyorsa bir maksimum marjdan söz edilemeyeceğinden yumuşak marj ifadesi ortaya atılmıştır. Bu durumda, hiperdüzlem ayrıldığında bazı örneklemler sınırın dışında kalabilmektedir ve marj δ hesaplanırken kesitin doğru bölümünde kalanlar için fark 0 olurken diğerleri için fark hesaplanmaktadır (Jiang, 2021).

Bir diğer destek vektör makinesi yöntemi ise lineer olmayan karar destek makinesi veya çekirdek hilesi olarak isimlendirilir. Özellikle örüntü tanıma problemleri gibi karmaşık problemlerin çözümünde lineer kesitler ile hiperdüzlemi ayırmak mümkün değildir. Bu durumda bir uzaydaki problemi daha yüksek boyutlu uzaya taşıyarak ilk uzayda lineer olarak ayrılamayan durumları daha sonra lineer olarak ayrılabilir hale getirme yöntemi izlenmektedir (Jiang, 2021).

Vektör nicemleme: Linear vector quantization olarak isimlendirilen bu yöntem en yakın komşu problemlerinin çözümündeki kompleksiteyi azaltmak amacıyla önerilmektedir. Bu yöntemde geliştirme veri kümesi içerisinden seçilen temsili değişkenler (prototipler) kullanılarak bunlar üzerinden farklılıklar hesaplanmaktadır. Sezgisel yöntemlerle her $i = 1,2,\dots,c$ sınıfından bir prototip $v_i \in R^p$ belirlenir ve x değişken vektörü aşağıdaki koşul sağlandığı durumda bu sınıfa atanır (Runkler, 2016):

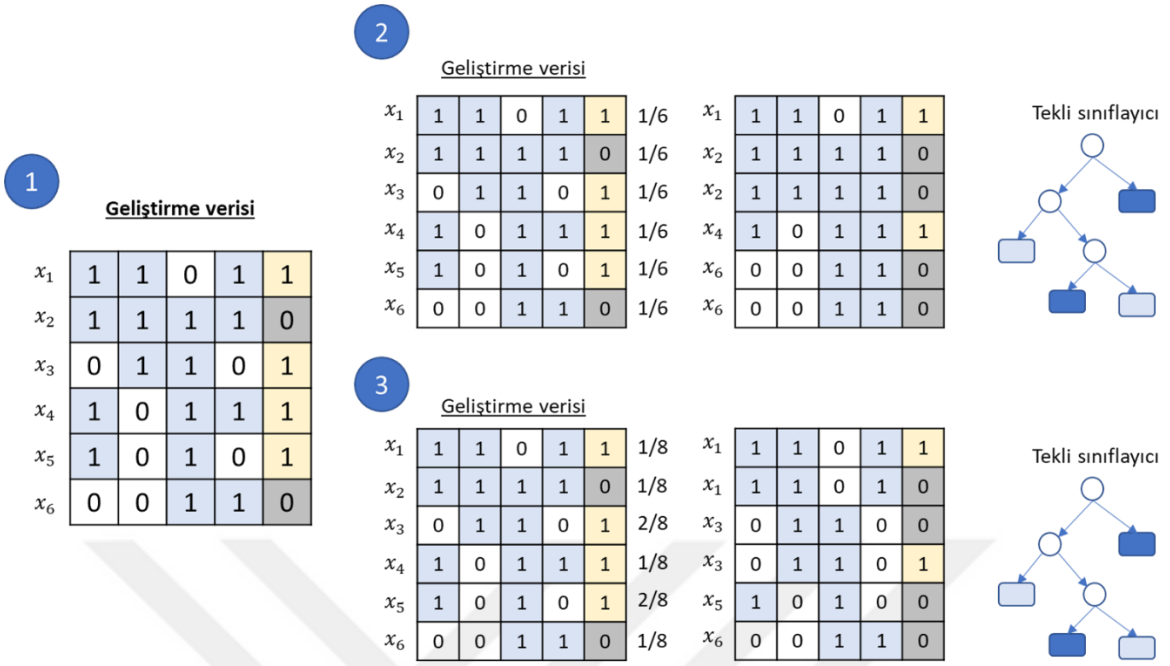
$$\|x - v_i\| = \min_{j=1,2,\dots,c} \|x - v_j\|$$

Vektör nicemleme yönteminde prototipler rassal şekilde seçilir. Seçilen ilk prototipten sonra seçilen diğer prototip eğer aynı sınıftan ise dahil edilir ve farklı bir sınıftan ise hariç tutulur. Bir

noktadan sonra yakınsama olacağından adım sayısı geliştirim sırasında azalmaktadır. Geliştirim veri kümesi üzerinden birkaç defa geçilerek işlemi sonlandırmak için gerekli kriter bulunmaktadır.

Artırma: Artırma (boosting) yöntemlerinin en popüler olanları ekstrem gradyan artırma (XgBoost) ve uyarlanabilir gradyan artırma (AdaBoost) algoritmalarıdır. Bu yöntemlerin analojisi şu çok basit öğrenme örneğiyle açıklanabilir. Bir sınıftaki tüm öğrencilerin yapılan testlerdeki tüm alanlarda başarılı olmasını istediğimizi varsayalım, ancak bazı öğrenciler birtakım derslerde başarısız olabilirler. Bu öğrencilerin tespit edilerek başarısız oldukları konularda eğitilmesi gerekir ki diğer öğrenciler ile eşit seviyeye gelebilsinler. Bu sebeple, bu öğrencilere bu dersler için özel sınıflar koyulabilir, ek zaman ayırmaları istenebilir ve böylece öğrencilerin başarı seviyesi artırılır. Aynı örneği makine öğrenmesi alanına uyarlarsak; her makine öğrenmesi öğrenicisi bir alt kümede eğitilir ve eğer bir öğrenicinin performansı düşük ise ona daha fazla ağırlık verilir. Buna artırma (boosting) yöntemi ismi verilir (Kumar, A., & Jain, M., 2020).

Artırma yönteminde, spesifik olarak üretilen alt örneklem kümeleri üzerinde uygulanan birçok zayıf sınıflayıcılar oylama yöntemi ile birleştirilmektedir. Bu örneklem kümeler geri iadeli örneklem şeklindedir ve n elemanlı bir öğrenme kümesinden n adet eleman rassal olarak seçilmektedir. İlk adımda her bir elemanın seçilme ihtimali 1/n'dir. Sonrasında her bir elemana farklı ağırlıklar da uygulandığından seçilme ihtimali değişmektedir. Ağırlıklandırma da hatalı tahminlenen kayıtlara daha yüksek ağırlık atama şeklinde yapılmaktadır ve dolayısıyla bu kayıtların seçilme ihtimali de daha yüksek olmaktadır. Her bir adımda da önceki adım sonucunda da hatalara göre yeniden ağırlıklandırma yapılmaktadır. Bu süreç aşağıdaki görselde resmedilmiştir (Kozak, 2019).



Şekil 16. Artırma Algoritmalarının Akışı

AdaBoost uygulamasında objelerin öğrenme kümesindeki ağırlıklarının bir önceki adımdaki hatalara göre belirlenmesi süreci aşağıdaki şekilde formüle edilmiştir. Burada we_i, x_i objesinin ağırlığıdır ve ${c_i}^j$ ise analiz edilen objenin sınıfıdır (Kozak, 2019).

$$\varepsilon(j) = \sum_{x_i} we_i[c_i \neq {c_i}^j]$$

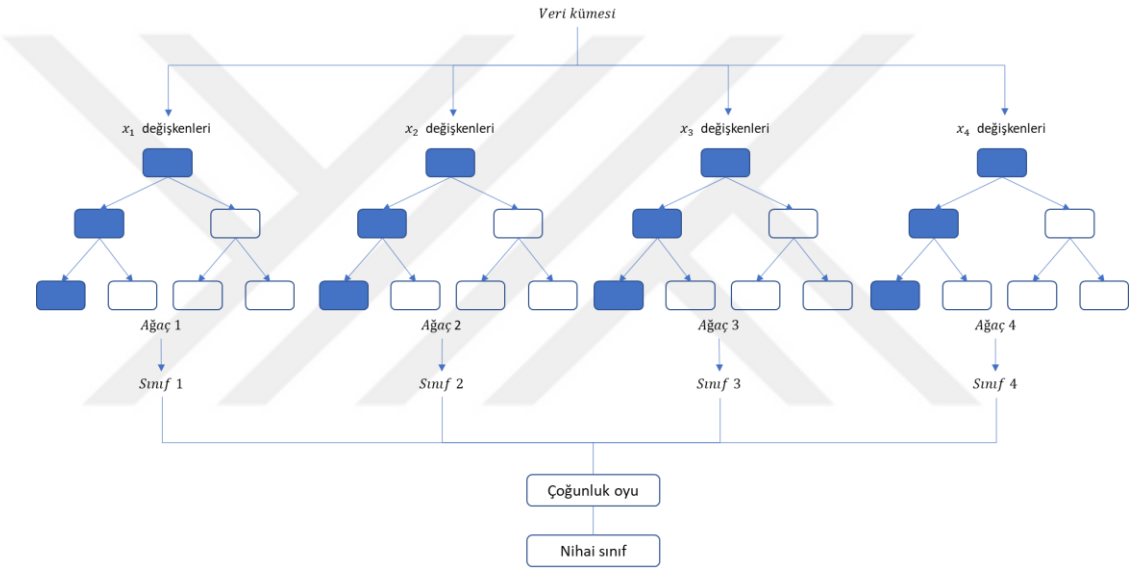
Sınıflandırma hatası 0.5'ten büyükse ağırlıklar katsayı ile çarpılmakta ve normalize edilmektedir (Kozak, 2019).

$$K(j) = \frac{1 - \varepsilon(j)}{\varepsilon(j)}$$

$$dDF(x) := \arg\max_{y \in Y} \sum_{j=1}^{J} \left(log \frac{1}{K_j} \right) h_j(x, y)$$

AdaBoost artırma algoritmalarında tüm öğrenme iterasyonları sıralı yapılmakla birlikte XGBoost algoritmasında ise paralel yapılmaktadır. XGBoost, model karmaşıklığını kontrol ederek ve aşırı öğrenmeyi azaltarak daha iyi bir model performansı sağlar. Sürekli değişkenlerde değişkenin bölüneceği noktanın belirlenmesi için bir algoritma kullanır (Quinto, 2020).

Rassal ağaç: Rassal ağaç algoritması sınıflandırma ve regresyon problemleri için bir seri karar ağacı kullanan bir algoritma türüdür. Aynı anda hem yanlılığı düşük tutup hem de varyansı azaltmak için torbalama (bagging) yöntemi kullanılır. Torbalama yönteminde model geliştirim verisinin alt örneklemleri üzerinden tekil ağaçlar oluşturulur. Bunun yanında bir de değişken torbalama yöntemi uygulanarak değişkenlerin de bir alt kümesi oluşturularak kullanılır. Değişken torbalama ile karar ağaçları arasındaki korelasyon azaltılmaya çalışılır ve uygulanmadığı takdirde karar ağaçlarının birbiri ile çok benzer olma riski vardır. Sınıflama probleminde çıktıların çoğunluk oylama sonucuna göre nihai model tahmini oluşturulur. Rassal ağacın çalışma yöntemi aşağıdaki şekilde görselleştirilmiştir (Quinto, 2020).

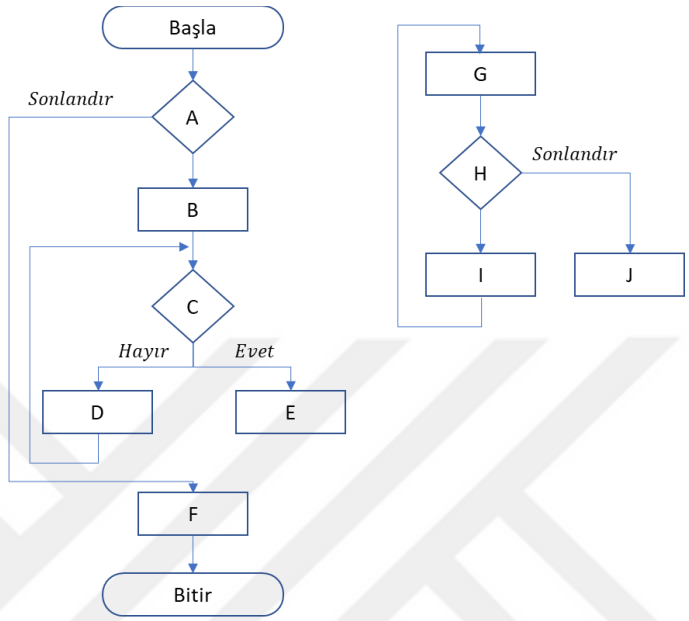


Şekil 17. Rassal Ağaç Algoritması

Rassal ağaç algoritmasının oluşturulma diyagramı aşağıdaki şekilde görselleştirilmiştir ve her adımın detayı da şu şekildedir (Kozak, 2019):

- A: Her karar ağacı için
- B: Alt örneklemi oluştur
- C: Ağaç tamamlandı mı?
- E: Tutarlılık oranını hesapla
- D: Bir sonraki örneklemi oluştur
- F: Karar ormanını oluştur
- G: Alt değişken kümesini seç
- H: Alt örneklemler ve alt değişken kümesi için

- I: Bölünme kural değerini belirle
- J: En iyi bölünmeyi seç

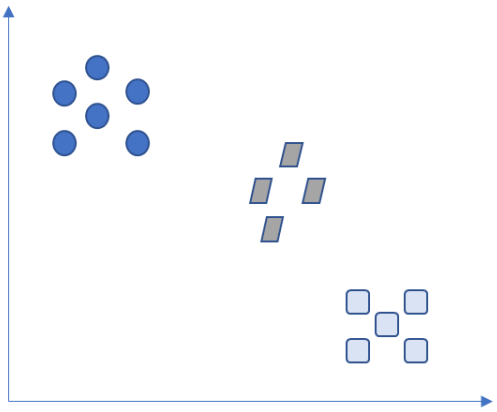


Şekil 18. Rassal Ağaç Algoritma Akışı

Karar ağaçlarını belirli alt örneklemler üzerinden geliştirmek tüm örneklemi ve tüm değişkenleri kullanmaya kıyasla hız kazandırmaktadır. Bu sebeple rassal ağaç algoritması karmaşık problemlerin çözümünde tercih edilebilen bir algoritmadır (Kozak, 2019).

2.4.2. Gözetimsiz Makine Öğrenmesi

Gözetimsiz öğrenme, bir diğer önemli makine öğrenmesi kategorisidir ve iş problemlerinin çözümünde gözetimsiz öğrenme de sıkça kullanılmaktadır. Gözetimli ve gözetimsiz öğrenme benzer veri kümeleri kullanılarak çalışmakla birlikte ikisi arasındaki temel fark gözetimsiz öğrenmede hedef değişkenin bulunmamasıdır. Yani, model doğru bir değeri tahmin etmek için geliştirilmemektedir. Gözetimsiz öğrenmede makine veri içerisindeki gizli örüntüleri ve benzerlikleri tespit etme ve bunu daha sonra uygulamada kullanma amacı taşımaktadır. Aşağıdaki şekilde gözetimsiz makine öğrenmesi görselleştirilmiştir (Singh, 2021).



Şekil 19. Gözetimsiz Makine Öğrenmesi

Bu yöntemin en önemli uygulama alanı müşteri verisini kullanarak müşterileri segmentlere ayırmaktır. Böylece benzer özellikli müşteriler aynı grupta yer alırken diğer gruptaki müşterilere göre de farklı sınıflanmaktadır. Validasyon perspektifinden bakıldığında, gözetimli makine öğrenmesinde bir performans kriteri bulunurken gözetimsiz makine öğrenmesinde bir performans kriteri de bulunmamaktadır. En sık kullanılan gözetimsiz makine öğrenmesi algoritmaları şunlardır (Singh, 2021):

- K-means sınıflandırma
- En yakın komşu haritalandırma

K-means algoritması, veriyi birbiri ile çakışmayan k adet sınıfa bölmeye çalışan basit bir gözetimsiz öğrenme algoritmasıdır. Her bir sınıf, sınıfın merkez noktası ile temsil edilir. K-means algoritmasının alışma mantığı şu şekildedir (Wu, 2012):

- Kullanıcı tarafından kaç adet sınıf oluşturulacağı yani k sayısı belirlenir.
- Veri içerisindeki her nokta en yakın merkez noktasına atanır.
- Her sınıfın merkez noktası atanan yeni noktalara göre güncellenir.
- Yeni noktaların atanması ile merkez noktası artık değişmeyene kadar bu işleme devam edilir.

2.5. Çeşitli Makine Öğrenmesi Konseptleri

2.5.1. Hiperparametre Optimizasyonu

Makine öğrenmesi problemlerinde ilgilenilmesi gereken iki tür değişken grubu vardır: parametreler ve hiperparametreler. Parametreler, kullanıcının çalıştığı veri kümesine göre doğrudan

algoritma tarafından belirlenen ve kullanıcının değişiklik yapması gerekmeyen değişkenlerdir. Hiperparametreler ise daha yüksek seviye parametreler olup model geliştirme öncesinde kullanıcının manuel olarak belirlemesi gereken değişkenlerdir ve verinin karakteristiğine veya algoritmanın öğrenme kapasitesine göre değişkenlik gösterir (Agrawal, 2021). Farklı algoritmaların hiperparametreleri şöyledir (Agrawal, 2021; Masood, 2021):

K- en yakın komşu algoritması için ayarlanması gereken hiperparametreler:

- K-en yakın komşu sayısı
- Ağırlıklar
- İndeksleme algoritması
- Uzaklık metriği (Öklidyen, Manhattan, Minkowski gibi)

Karar vektör makineleri algoritması için ayarlanması gereken hiperparametreler:

- Kernel (çekirdek)
- C düzenleştirme parametresi
- Gamma düzenleştirme parametresi
- Polinom derecesi

Karar ağacı algoritması için ayarlanması gereken hiperparametreler:

- Algoritma
- Ağacın derinliği
- Her kırılımdaki minimum kayıt sayısı
- Minimum yaprak sayısı

XGBoost algoritması için ayarlanması gereken hiperparametreler:

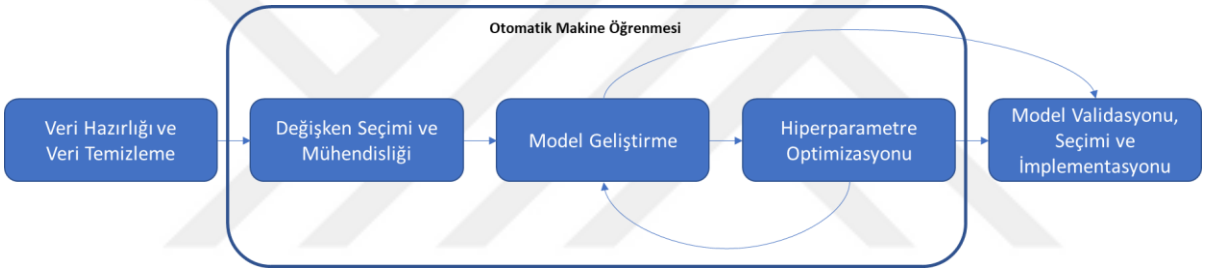
- Alt örneklem oranı
- Colsample_bytree: Kolonların her bir ağaçta seçilme oranı
- Colsample_bylevel: Kolonların her bir seviyede seçilme oranı
- Gamma düzenleştirme parametresi
- Booster: gbtree veya gblinear
- Öğrenme oranı
- Maksimum derinlik
- N_estimators: ağaç sayısı

- Reg_alpha:L1 düzenleştirme parametresi
- Reg_lambda: L2 düzenleştirme parametresi

2.5.2. AutoML: Otomatik Makine Öğrenmesi

Daha önceki bölümlerde veri analitiği süreci özetlenmiştir. Bu süreç içerisinde veri hazırlığı, veri ön işlemleri, model seçimi, hiperparametre optimizasyonu ve bunun sonucunda en optimal modelin seçimi şeklinde adımlar yer almaktadır. Ancak son yıllarda bu süreçlerin uçtan uca otomatize edildiği otomatik makine öğrenmesi (AutoML) akışları ve platformları da hayatımıza girmiştir (Masood, 2021).

Otomatik makine öğrenmesi ile hedef, model geliştirme adımlarını basitleştirmek ve demokratikleştirmektir. Geçmişte özellikle model seçimi ve hiperparametre optimizasyonu konularına ağırlık verilmiştir. Ancak günümüzde, tüm sürecin otomatize edilmesi hedeflenmektedir ve bu da aşağıdaki görselde gösterilmektedir (Masood, 2021).



Şekil 20. AutoML - Otomatik Makine Öğrenmesi

2.5.3. Açıklanabilirlik

Makine öğrenmesi modellerinin başarılı bir şekilde implementasyonu ve kullanımı modellerin açıklanabilirliğine oldukça bağlıdır ve sektördeki son kullanıcılar açıklanabilir olduğu için daha basit algoritmaları veya daha düşük performanslı modelleri daha karmaşık modellere tercih edebilmektedir. Bu sebeple, modeller farklı geçmişten gelen veya farklı bilgi seviyesinde olan kişilerin anlamasına olanak vermelidir (Jung, 2022).

Makine öğrenmesi modellerinin açıklanabilir olması sadece anlaşılması değil ayrıca hukuki veya etik sebeplerle de gereklidir. Modeller ile verilen kararların denetimi noktasında veya her kararın eşitlik/adalet ilkelerine uygunluğu noktasında da açıklanabilirlik zorunludur. Makine öğrenmesi modellerinin açıklanabilirliği ile birlikte hayatımıza giren konsept XAI yani açıklanabilir yapay zeka olmuştur. Bununla birlikte kara kutu olarak isimlendirilen makine öğrenmesi modelleri beyaz kutu haline getirilmektedir. Bu amaçla en sık kullanılan algoritma SHAP değeridir. SHAP değeri oyun teorisi için geliştirilmiştir ve oyun içerisinde her oyuncunun marjinal etkisini bulmayı amaçlamaktadır. Makine

öğrenmesi modellerinde de her değişkenin nihai çıktıya marjinal etkisini belirlemek için bu yaklaşımdan faydalanılmaktadır (Rothman, 2020).

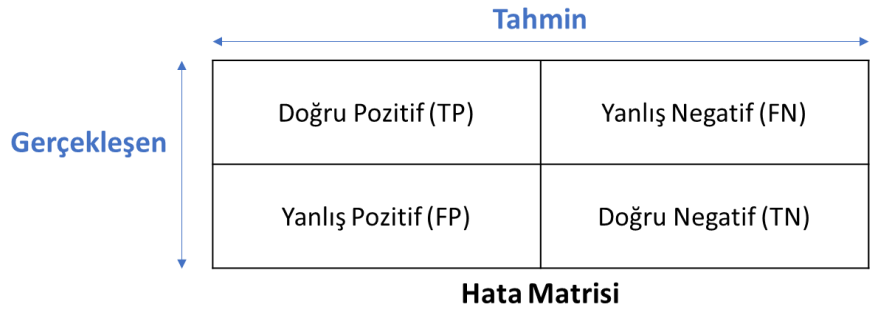
2.5.4. Model Performans Metrikleri

Sınıflandırma problemlerinde model performans metrikleri aşağıdaki şekildedir ve tanımlamaları kısaca özetlenmiştir.

Sınıflandırma problemlerinde, model performansı özetle gerçekleşen değer ile tahminlenen değer arasındaki uyumun ne derece yüksek olduğunu hesaplamak üzerinedir. Bu hesaplamaların da çoğu aşağıda verilen "Hata Matrisi" üzerinden çıkarılabilmektedir. Hata matrisinin TP ve TN hücreleri doğru şekilde tahmin edilmiş kayıtları göstermektedir. Örneğin, insan kaynaklarında işten ayrılma problemi için:

- TP: işten ayrılan ve ayrılacağı tahmin edilmiş olan kayıtlar
- TN: işten ayrılmayan ve ayrılmayacağı tahmin edilmiş olan kayıtlar
- FP: işten ayrılmayan ancak ayrılacağı tahmin edilen kayıtlar (Tip 1 Hata)
- FN: işten ayrılan ancak ayrılmayacağı tahmin edilen kayıtlar (Tip 2 Hata)

Tip 1 hatanın yüksek olması sektörde yanlış alarm yaratılmasına ve işten ayrılmayacağı halde yersiz önlem alınmasına neden olabilir. Tip 2 hatanın yüksek olması ise çalışanın işten ayrılması önceden bilinemediği için hiç önlem alınmamasına ve bu sebeple de geç kalınma problemine yol açabilir. İnsan kaynakları sektörü dışında finans sektöründeki dolandırıcılık/suiistimal örneği de verilebilir. Bir müşterinin dolandırıcı olduğu halde bunun tespit edilememesi "yanlış negatif" örneğidir ve finansal kurum için maddi kayba ve itibar kaybına yol açar. Diğer yandan ise, dolandırıcı olmadığı halde dolandırıcı olduğu düşünülen bir müşterinin alışverişi bloke edilir veya hesabı dondurulursa da müşteri memnuniyeti kaybına ve sonunda müşteri kaybına sebep olur. Bu örneklerden de anlaşılacağı üzere, Tip 1 ve Tip 2 hataların dengesi model performansı için kritiktir.



Şekil 21. Hata Matrisi

Doğruluk: Doğruluk, modeldeki doğru tahmin ettiğimiz kayıtların toplam içerisinde oranıdır.

$$Doğruluk = \frac{TP + TN}{TP + FP + TN + FN}$$

Kesinlik: Kesinlik, pozitif olarak tahmin edilen kayıtların kaç tanesinin pozitif olduğu bilgisidir.

$$Kesinlik = \frac{TP}{TP + FP}$$

Duyarlılık: Duyarlılık, pozitif olarak tahmin edilmesi gereken kayıtların kaç tanesinin pozitif olarak tahmin edildiği bilgisidir.

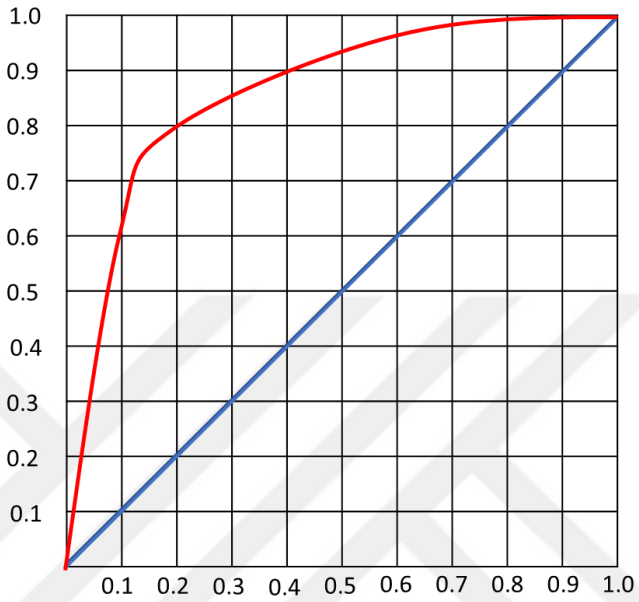
$$Duyarlılık = \frac{TP}{TP + FN}$$

F1 Skoru: Bu skor, kesinlik ve duyarlılık metriklerinin harmonik ortalaması olup kombine bir hesaplamadır.

$$F1\ Skoru = 2 * \frac{Kesinlik * Duyarlılık}{Kesinlik + Duyarlılık}$$

Gini ve AUC: ROC eğrisi model performansını yorumlamak için en çok kullanılan grafiklerden biridir. Aşağıda örneği verilen bu eğride, yatay eksen "yanlış pozitif" oranını ve dikey eksen ise "doğru pozitif" oranını göstermektedir. Dolayısıyla, eğri ne kadar yüksek ise doğru pozitif

oranı yanlış pozitif oranına göre o derecede baskın demektir ve o derecede model performansı yüksektir. En iyi model, bu kırmızı eğrinin köşeli olduğu ve tüm kayıtların doğru pozitif olduğu durumda ortaya çıkmaktadır. En kötü model ise kırmızı eğrinin köşegen üzerinde olduğu ve kayıtların yarısının doğru pozitif, diğer yarısının ise yanlış pozitif olduğu durumdur.



Şekil 22. ROC Eğrisi

AUC (Area Under Curve=Eğri Altındaki Alan), kırmızı eğri altında kalan alanın büyüklüğü olup en iyi modelde 1'dir. Rassal bir modelde ise bu alan 45 derecelik köşegenin altında kalan alan olacağından 0.5'dir.

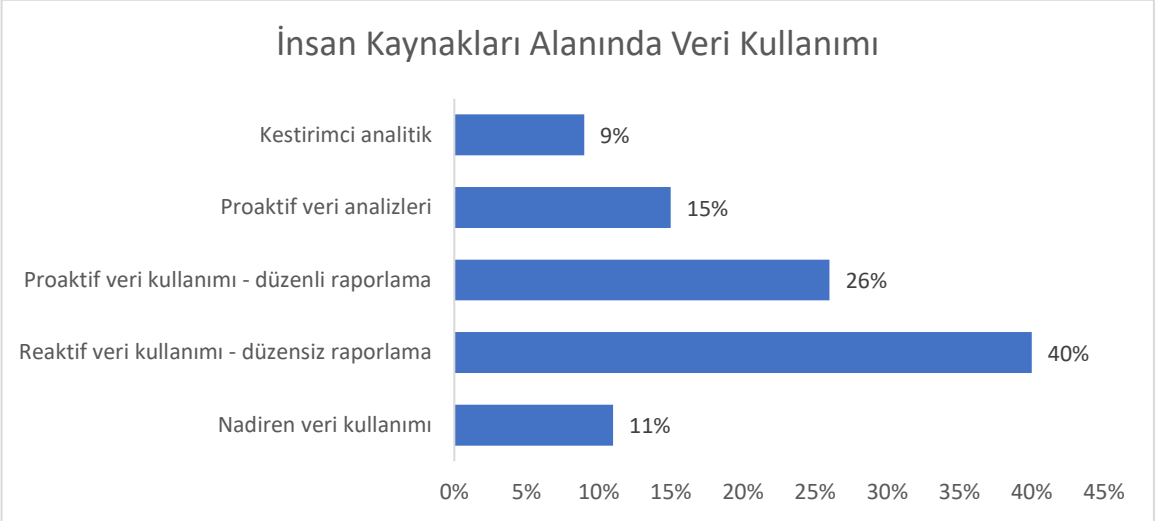
Gini ise 2*AUC-1 olup 0 ile 1 arasında değişen bir metriktir. En iyi model 1 Gini performansına sahip iken en kötü model ise 0 performansına sahiptir.

3. İNSAN KAYNAKLARI ANALİTİĞİ

Bir önceki bölümde veri analitiği, büyük veri, makine öğrenmesi ve yapay zeka kavramları detaylı şekilde açıklanmıştır. Bu bölüm ise bu uygulamaları insan kaynakları alanındaki yansımasını özetlemektedir.

3.1. İnsan Kaynakları Verisi

CIPD tarafından yapılan bir çalışmada şirket yöneticilerinin insan kaynakları yönetiminde veriden ne ölçüde faydalandıkları sorulmuş ve katılımcıların büyük çoğunluğu çeşitli düzeylerde veri kullanmakta olduğunu ifade ederken sadece %9'unun kestirimci analitik çalışmalar yaptığı ortaya çıkmıştır. Bu soruya verilen cevapların genel dağılımı şu şekildedir:



Şekil 23. İnsan Kaynakları Alanında Veri Kullanımı

Bu bölümde insan kaynakları alanında hangi veri ve metriklerin kullanıldığı, bu verilerin sahipliği ve insan kaynakları alanında veri kullanımında karşılaşılan ana sorunlar anlatılmaktadır.

3.2. İnsan Kaynakları Ham Veri Kaynakları

Visier tarafından yapılan bir saha araştırmasında insan kaynakları profesyonellerinin kurumlarında hangi veri kaynaklarını kullandıkları sorulmuş olup gelişmiş ve gelişmekte olan şirketler kırılımında bu veri türlerine sahiplik incelenmiştir. Aşağıdaki tablo, veri türlerine sahipliği özetlemektedir (Visier, 2020).

Tablo 4. İnsan Kaynakları Alanındaki Veri Kaynakları

Veri Türü	Gelişmiş Kurum	Gelişmekte Olan Kurum
Temel insan kaynakları verisi	%81	%88
Yetenek yönetimi verisi (performans, öğrenme, işe alım)	%85	%65
Çalışan memnuniyeti verisi	%63	%35
İK spesifik veri ambarı	%54	%32
İşgücü yönetimi verisi (iş saatleri, devamsızlık, izinler, vb.)	%37	%33
Dışsal karşılaştırma verileri (ücretlendirme, çalışan işten ayrılması, vb.)	%37	%33
Finansallar	%40	%27
Genel veri ambarı	%27	%13
Operasyonel sistem	%23	%13
Satış/Müşteri İlişki Yönetimi verisi	%13	%6
Güvenlik/teftiş verisi	%6	%8
Pazarlama verisi	%10	%2

3.2.1. İnsan Kaynakları Performans Metrikleri

İnsan kaynakları alanındaki ham veriden hareketle üretilen ana performans metrikleri aşağıdaki üç alt kırılımda incelenebilir ve bu kırılımların altında yer alan metrikler için örnekler Fitz-enz, J. ve Mattox, J. R. II tarafından yazılan kitap baz alınarak örnekler çoğaltılmıştır (2014):

- Etkinlik metrikleri
- Etkililik metrikleri
- İş çıktısı metrikleri

Etkinlik metrikleri, insan kaynakları departmanının yaptığı çalışmaların performansını ölçen metriklerdir. Bu çalışmalar genellikle işe alım, ücret belirleme, performans değerlendirme alanlarını kapsamaktadır. Etkinlik metrikleri ile ilgili örnek bir tablo aşağıda sunulmuştur:

Tablo 5. İnsan Kaynakları Etkinlik Metrikleri

Metrik	Veri Tipi	Geçen Yıl Gerçekleşen	Bu Yıl Planlanan	Bu Yıl Gerçekleşen	Başarı Ölçümü
Doldurulan açık pozisyonlar / ay	Adet				
İlgili Ay doldurulan açık pozisyonlar / Tüm açık pozisyonlar	Yüzde				
Ortalama açık posizyon doldurma süresi	Gün sayısı				
Sunulan ortalama maaş tutarı	Parasal değer				
Ortalama işe alım maliyeti	Parasal değer				
Performans değerlendirme süresi	Gün sayısı				
Performans değerlendirme maliyeti	Parasal değer				

Etkililik metrikleri, insan kaynakları departmanının yaptığı çalışmaların ne kadar etkili olduğunu ve esas amaca ne kadar hizmet ettiğini ölçen metriklerdir. Etkililik metrikleri ile ilgili örnek bir tablo aşağıda sunulmuştur:

Tablo 6. İnsan Kaynakları Etkililik Metrikleri

Metrik	Veri Tipi	Geçen Yıl Gerçekleşen	Bu Yıl Planlanan	Bu Yıl Gerçekleşen	Başarı Ölçümü
Yeni işe alınan kaynakların 90 gün sonundaki ilk performans notları	Likert ölçeği ile 1-9				
Yeni işe alınan kaynakların 1 yıl sonundaki performans notları	Likert ölçeği ile 1-9				

Çalışan anketlerinin sonuçları	Ortalama skor				
Yeni işe alınan çalışanın yetkin seviyeye ulaşma süresi	Ay sayısı				
İç müşteri anket sonuçları	Ortalama skor				
Çıkış mülakatı değerlendirme sonuçları	Ortalama skor				

İş çıktısı metrikleri, insan kaynakları departmanının yaptığı çalışmaların kurumun nihai hedeflerine ne derecede katkı sağladığını ölçen metriklerdir. İş çıktısı metrikleri ile ilgili örnek bir tablo aşağıda sunulmuştur:

Tablo 7. İnsan Kaynakları İş Çıktısı Metrikleri

Metrik	Veri Tipi	Geçen Yıl Gerçekleşen	Bu Yıl Planlanan	Bu Yıl Gerçekleşen	Başarı Ölçümü
Çalışan memnuniyeti anket sonuçları	Ortalama skor				
Verimlilik	Ortalama çalışan başına fatura edilebilir süre				
90 gün içinde işten ayrılma oranı	Yüzde				
1 yıl içinde işten ayrılma oranı	Yüzde				
Gelir	Gelir tutarı				
Karlılık	Kar tutarı				
Yeni geliştirilen inovatif ürün/hizmet sayısı	Adet				

3.2.2. İnsan Kaynakları Verisinin Sahipliği

İnsan kaynakları alanındaki veriye dayalı çalışmaların zor olmasının en önemli sebeplerinden biri veri sahipliğinin farklı departmanlarda olmasıdır. Bu da bütüncül bir bakış açısıyla personele dair tüm verinin toplanmasını ve işlenmesini zorlaştırmaktadır.

İnsan kaynakları performans yönetiminde kullanılacak veri alanları ve bunların sahipliği için örnek bir tablo aşağıda paylaşılmıştır (Fitz-enz, J. & Mattox, J. R. II, 2014):

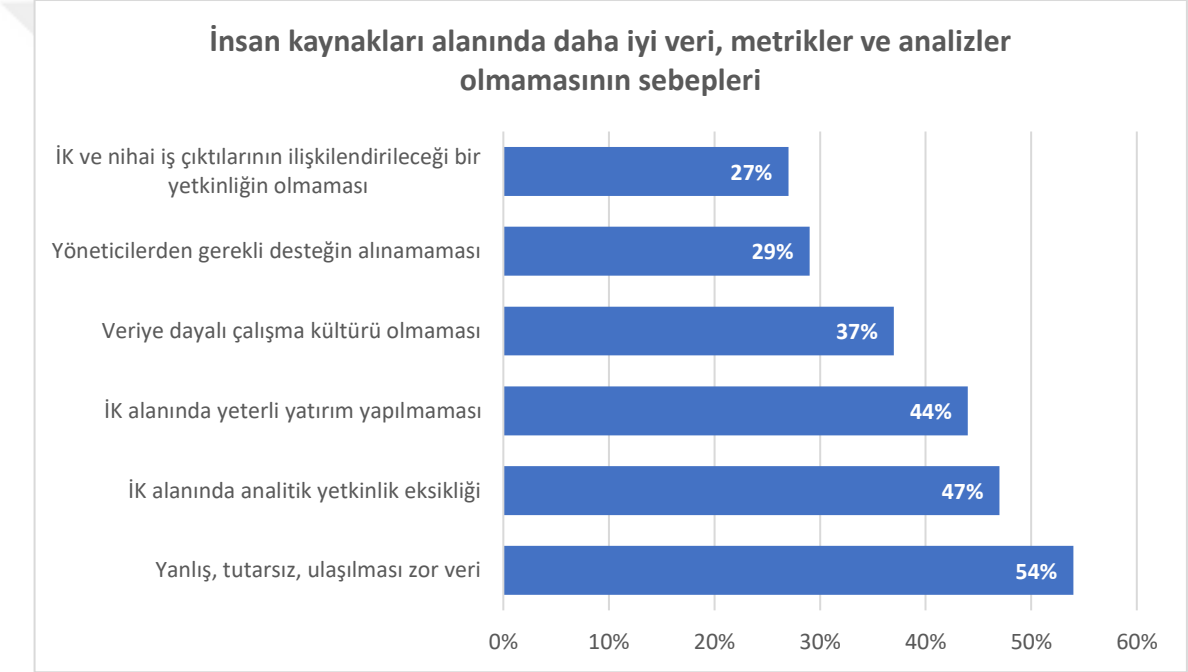
Tablo 8. İnsan Kaynakları Verisinin Sahipliği

Veri Sahibi Departman	Etkinlik Metrikleri	Etkililik Metrikleri	İş Çıktısı Metrikleri
İnsan Kaynakları	• Doldurulan açık pozisyonlar / ay • İlgili Ay doldurulan açık pozisyonlar / Tüm açık pozisyonlar • Ortalama açık pozisyon doldurma süresi • Sunulan ortalama maaş tutarı • Performans değerlendirme süresi	• Yeni işe alınan kaynakların performans notları • Çalışan anketlerinin sonuçları • Yeni işe alınan çalışanın yetkin seviyeye ulaşma süresi • İç müşteri anket sonuçları • Çıkış mülakatı değerlendirme sonuçları	• Çalışan memnuniyeti anket sonuçları • 90 gün içinde işten ayrılma oranı • Verimlilik
Finans	• Ortalama işe alım maliyeti • Yeni işe alımların eğitim maliyeti • Performans değerlendirme maliyeti		• Gelir • Karlılık
Satış			• Yeni geliştirilen inovatif ürün/hizmet sayısı • Çalışan başına ortalama satış adedi • Yeni müşteri kazanımı
Kalite			• Ürün veya hizmetteki hata oranı

Müşteri İlişkileri Yönetimi			• Müşteri sadakati • Müşterilerin çalışan memnuniyeti anketi

3.2.3. İnsan Kaynakları Verisi Kullanımında Karşılaşılan Sorunlar

Harvard Business Review ve Visier tarafından yapılan bir araştırmada insan kaynakları ve yetenek yönetimi alanında çalışan profesyonellere insan kaynakları alanında daha iyi veri, metrikler ve analizler olmamasının sebepleri şu şekilde sıralanmıştır (Harvard Business Review Analytic Services, 2014):



Şekil 24. İnsan kaynakları alanında anket sonuçları

3.3. İnsan Kaynakları Analitiği Tarihçesi

İnsan kaynakları analitiği ne kadar yeni bir kavram gibi görünse de temellerinin çok daha önceden atıldığı söylenebilir.

1911'de makine mühendisi olan ve endüstri mühendisliğinin de kurucusu olarak kabul edilen Frederick Taylor, çalışanların performansını ölçmek ve onların endüstriyel verimliliklerini artırmak amacıyla yaptığı ve "metot etüdü" adını verdiği araştırmaları ile çalışanların verimliliğini ölçmüştür. Bu çalışmalarını ise "Bilimsel Yönetimin Prensipleri" adını verdiği kitabında toplamıştır. İşçilerin

yapacakları işe göre seçimi ve eğitimi, iş bölümü alanları yanında işçi-işveren ilişkileri, planlama ve kontrol fonksiyonlarının da işçiler tarafından değil de yönetim tarafından yapılması gerektiğini öne sürmüş ve fiziksel ve mental işleri birbirinden ayırmıştır (Tanyaş, M., 2000).

Taylor, prensiplerini kanıtlamak için "Pik Demir Benzetimi"ni kullanmış ve iş ve zaman etüdü çalışmalarını işçilerin en etkili ve mutlu çalışacakları optimum iş yükünü belirlemek amacıyla kullanmıştır (Govekar, P., L. & Govekar, M. A., 2012). Taylor'ın çalışmaları verimliliğin altında yatan veriyi analiz etme konusunda yol gösterici olmuştur.

E. L. Thorndike 1920'de sosyal zekaya yönelik ilk psikometrik görüşleri paylaşmış ve zekayı soyut zeka, mekanik zeka ve sosyal zeka olarak üçe ayırmıştır. Soyut zeka fikirlerin algılanması ve yönetilmesi yeteneği, mekanik zeka ise fiziksel objelerin manipüle edilmesi yeteneği olarak tanımlanmıştır. Sosyal zeka ise insan ilişkilerinde akıllıcı davranmak olarak tanımlanmıştır. 1921 ve 1923'te Carl Gustave Jung ise sosyal zekayı çalışma konusuna bağlamış ve kişisel yeteneklerin yanı sıra kişiliğin ve sosyal zekanın da verimliliği etkilediği sonucunu çıkarmıştır (Kihlstrom, J., F. & Cantor, N., 2000).

1978'de Dr. Jac Fitz-enz, insan kaynakları metriklerini açıklamış ve Saratoga Enstitüsü ve şimdiki ismiyle İnsan Kaynakları Yönetimi Topluluğu (SHRM) ile birlikte bunları 1984'te bir kitapta toplamıştır. Bu metriklerden bazıları şimdilerde kestirimci analitik konularına konu olan gönüllü veya gönülsüz işten ayrılma, açık pozisyonu doldurma süresi, işe başlama süresi, çalışan başına elde edilen kar gibi metriklerdir (Isson & Harriott, 2016).

1996'da Kaplan ve Norton tarafından geliştirilen "Balanced Scorecard" konseptinden yola çıkarak 2001 yılındaki kitapları "The HR Scorecard" ile Becker, B. E. vd. insan kaynakları aktivitelerinin şirketin kurumsal strateji ve aktiviteleri ile uyumluluğunu sağlayan insan kaynakları skorkartlarının organizasyonel çıktılara katkısını göstermişlerdir. Bu çalışmada, şirket hedeflerinden yola çıkarak insan kaynakları hedefleri belirlenmekte, sonra insan kaynaklarının finans, müşteri, süreç vb. hedefleri operasyonel hedefler ile ilişkilendirilmektedir (Strohmeier, S. & Piazza, F., 2015)

2005'te ilk defa performans yönetim sistemi kullanılmıştır. Performans yönetim sistemleri, yetenek yönetimi, işe alım, öğrenme ve gelişim ile ücretlendirme gibi ana insan kaynakları süreçlerini otomatize etmekte ve geliştirmektedir. Bu sistemler, özgeçmiş, performans geçmişi, eğitim verisi gibi çok çeşitli insan kaynakları verilerini de saklamaktadır. Günümüzde bu sistemler içerisinde sosyal medya verisi ve çalışanın dijital ayak izi verisi gibi alternatif veri kümeleri de dahil edilmiş durumdadır (Isson & Harriott, 2016).

2010'lar itibariyle de insan kaynakları analitiği alanında verinin analiz edilmesi ve sonuç çıkarılması dışında kestirimci analitik modelleri ve büyük veri çalışmaları uygulanmaya başlanmıştır. Bunun yanı sıra teknolojinin gelişimi de insan kaynakları sürecini desteklerken yeni veri kaynaklarının kullanımında da ivme kazandırmıştır.

3.4. İnsan Kaynakları Analitiği Süreci

İnsan kaynakları analitiği süreci planlamadan değerlendirmeye kadar olan 9 adımdan oluşmaktadır (CIPD, 2020):

Adım 1 - Planlama: Analitik aktivitenin amacı ve hedefinin belirlenmesi adımıdır. Müşteri ihtiyaçlarının belirlenmesi ve bu ihtiyaçları karşılayacak analitik çalışmalar ile eşleştirilmesi bu adımda yapılmaktadır.

Adım 2 – Ana başarı faktörlerinin belirlenmesi: Projenin başarılı olduğunu gösterecek ölçütlerin belirlenmesi adımıdır. Bu ölçütlere örnek olarak zamanında teslimat, proje etkisi ve kullanıcı geri bildirimleri verilebilir.

Adım 3 – Veri testi: Mevcut olan verinin belirlenmesi ve bu verinin kalitesinin değerlendirilmesi adımıdır. Verideki eksikler ve bu eksiklerin nasıl giderileceği bu adımda çalışılmaktadır.

Adım 4 – Süreç tasarımı: Ekip üyelerinin rol ve hedeflerinin belirlenmesi adımıdır. Kaynak ihtiyaçlarının ve proje taraflarının belirlenmesi bu adımda yapılır.

Adım 5 – Veri toplama stratejisi tasarımı: Analitik çalışmanın veri toplama ve işleme aşamalarının tasarlandığı adımdır.

Adım 6 – Veri toplama: Veri kaynaklarından verinin toplanması adımıdır. Bu mevcut veri ambarlarından verinin toplanması olabileceği gibi anket vb. yöntemler ile var olmayan yeni bir veriyi toplamak şeklinde de olabilir.

Adım 7 – Veri analizi: Müşteri ihtiyaçlarına bağlı olarak verinin analiz edilmesi ve bundan içgörü yaratılması adımıdır. Bu istatistiksel analiz üç şekilde olabilir (Fitz-enz, J. & Mattox, J. R. II, 2014) (Gene, P., 2015):

- Betimsel istatistikler
- Tahminsel istatistikler

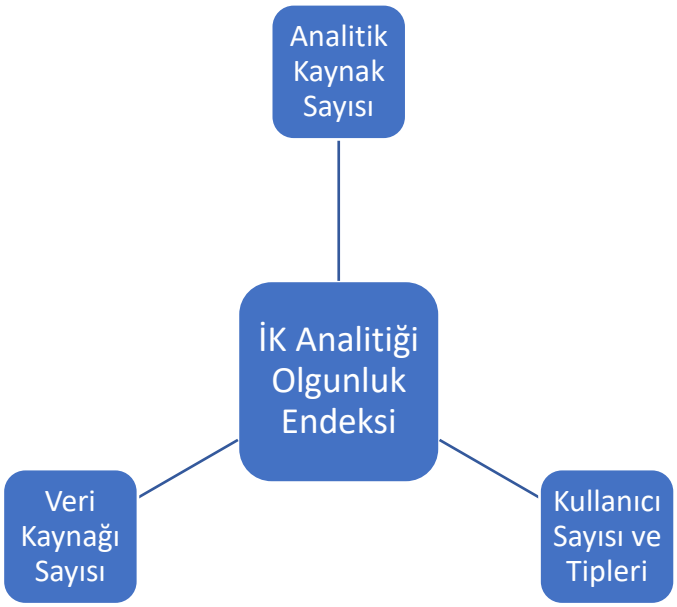
- Transformatif çalışmalar (Optimizasyon)

Adım 8 – Veri raporlaması: Düzenli ve basit bir şekilde probleme karşı bir çözüm sunma ve gelecek çalışmalar için gerekli yeni veri kümelerinin belirlenerek yönlendirme yapılması adımıdır.

Adım 9 – Değerlendirme: veri analitiği sürecinin değerlendirilmesi ve etkisinin ölçümlenmesi adımıdır. Bu değerlendirmeye göre sürecin en baştan alınması da söz konusu olabilmelidir.

3.5. İnsan Kaynakları Analitiği Olgunluk Endeksi Ölçümü

İnsan kaynakları analitiği olgunluk endeksi ölçümü (PAMI), süreç olgunluğu, İK analitiğinin nasıl çalıştığı, kullanıcı tipleri, karar vermede verinin önemi, veri kaynakları, başarı için ortaklıklar, değişim yönetiminin organizasyon içerisinde hangi derece ve kapsamda yapıldığı gibi çeşitli sorulara verilen cevaplardan elde edilen adet ve skorların toplamıdır. Olgunluk, İK analitiği çözümlerinin benimsenmesi ile sofistike bir yönetimin sağlanması anlamına gelmektedir. Ayrıca, kullanıcı sayısının artması ve daha fazla veri kaynağı kullanımı da organizasyonun iş çıktılarına etki edilmesine olanak sağlamaktadır. Değişim yönetimi kültürü ve veri ile kullanım olgunluğu uygulamaları da İK analitiği çözümlerinin kolay implemente edilmesine olanak sağlamaktadır.



Şekil 25. İnsan Kaynakları Analitiği Olgunluk Endeksi

3.6. İnsan Kaynakları Analitiği Sorumluluk Alanları

Şirketler için en önemli kaynaklardan biri çalışandır ve hammadde, malzeme, makina, teçhizat, enerji gibi işletmenin ve sunulan ürün veya hizmetin devamlılığı için vazgeçilmezdir. Bunun farkına varan işletmeler, son yıllarda çalışanların tüm ihtiyaçları ile ilgilenen alan olan İnsan Kaynakları alanına daha fazla odaklanmış ve Yetenek Yönetimi, İnsan Sermayesi Yönetimi gibi çeşitli yeni departmanlar ile bu fonksiyonlarını güçlendirmişlerdir. Fonksiyonun daha etkin çalışabilmesi için bu departmanın kullandığı araçlar arasında veri de en önemlilerden biri olarak yerini almıştır.

Isson ve Harriott (2016) ise İnsan Analitiği isimli kitapları ile iş analitiğinden insan analitiğine geçişi özetlemiş ve sonrasında da İnsan Kaynakları alanında birçok alanda analitik yöntemlerin teoride nasıl uygulanabileceğini açıklamıştır. Bu alanlar işgücü planlama, kaynak işe alımı, performans yönetimi, çalışan hayat boyu değerinin hesaplanması şeklindedir. Fitz-enz & Mattox (2014) tarafından yazılan İnsan Kaynakları'nda Kestirimci Analitik kitabında da analitik yöntemlerden çeşitli İnsan Kaynakları başlıklarında nasıl yararlanılabileceği anlatılmaktadır.

Bu bölümde, 2 numaralı başlık altında sözü edilen İnsan Kaynakları sorumluluk alanları analitik bir bakış açısı ile değerlendirilmekte ve veriye dayalı olarak yapılan çalışmalar özetlenmektedir.

3.6.1. İşgücü Planlama Analitiği

İşgücü planlama, basit tanımıyla doğru sayıda ve doğru nitelikteki çalışanları doğru işlere ve doğru zamanda alma işine verilen isimdir (Sinclair, 2004). Daha kapsamlı bir tanım ise şu şekilde yapılabilir: "bir organizasyonun işgücü ihtiyacı talebinin tahmin edilmesi ile bu talebi karşılayacak işgücü arzının boyutunun, doğasının ve kaynaklarının değerlendirilmesi" (Reilly, 1996).

Herhangi bir şirkette doğru bir işgücü planlaması olması için bu sürecin tüm elemanlarının verisinin doğru şekilde toplanması, analiz edilmesi ve raporlanması zorunludur. Bu süreç içerisinde gelecekteki işgücü büyüklüğünü sağlayabilmek için aşağıdaki operasyonlara dair bilgiler düzenli olarak toplanmalıdır (Schweyer, A., 2004):

1. İçeriden ve dışarıdan açık pozisyonlara başvuruların toplanması ve verilerin işlenebilir formatta saklanması
2. İşe alım sürecine dair testler, değerlendirmeler, çalışan referans sorguları, görüşme notları, vb. bilgilerin düzenli şekilde tutulması
3. Mevcut çalışanlara yönelik yetenekler, yetkinlikler, ilgi alanları gibi bilgileri içeren envanterlerin doldurulup saklanması

4. Performans görüşmesi ve çalışma arkadaşı değerlendirmeleri
5. Eğitim sonuçlarına ilişkin bilgilerin tutulması
6. Yetenek testlerinin, psikometrik testlerin, kişilik analizlerinin yapılması
7. Ödül ve takdir programlarına dair bilgiler

3.6.2. Çalışan Bulma Analitiği

İşgücü bir şirketin en önemli kaynaklarından birini oluşturmakta ve şirketin performansına büyük derecede katkı sağlamaktadır. Bu nedenle, şirketler doğru çalışanı doğru kanaldan ve doğru zamanda bulmak zorundadır. Günümüzde bu süreç profesyonel sosyal medya, büyük veri çözümleri ve bu işe özel uygulamalar sayesinde kolaylaşmıştır.

Veriye dayalı olarak piyasada iş arayışında olan çalışanların başvuracağı şekilde iş ilanlarının dizayn edilmesi, geçmişte paylaşılan ilanların internette açılma veya başvuru istatistiklerinin değerlendirilerek bunlardan ders çıkarılması iş ilanı yönetiminde yapılan uygulamalardan bazılarıdır. Bunun yanında, başvuruların değerlendirilmesi ya da başvurmayan kişilere doğrudan şirket tarafından teklif sunulması süreçlerinde de veri madenciliği ile özgeçmişler taranarak şirkete ve açık pozisyona uygun adaylar ile iletişime geçilmesi sağlanmaktadır (Isson & Harriott, 2016). Bunlar dışında marka yönetimi altında sosyal medya reklamlarında kişilerin arama geçmişlerine özgü çıkan reklamlar da şirketlerin aday havuzunda olarak nitelendirdiği çalışanların önüne direkt olarak çıkarılabilmekte ve bu kanallardan da adayların şirkete başvuru yapması sağlanabilmektedir. Deloitte tarafından yapılan bir çalışmada üst düzey yöneticilerin %62'si çalışan bulma ve reklam verme amacıyla sosyal platformları kullandığını açıklamıştır. Bu çalışmada, şirketler ayrıca analitik alanından faydalanmaya başladıklarını söylerken %54'ü bu alanda kendini "zayıf" durumda olarak tanımlamıştır (Bohdal-Spiegelhoff, 2014).

3.6.3. Çalışan İşe Alım Analitiği

Çalışan işe alım süreci çalışanların ilana başvurmasından işe alımına kadar olan süreci kapsamaktadır. Dolayısıyla, ilana yapılan başvurular arasından sınırlı sayıda ve uygun başvurunun seçilmesi, bunların değerlendirilmesi, ön görüşmelerin ve mülakatların yapılması ile işe alım kararının verilmesi adımları bu kapsama girmektedir.

CIPD tarafından her yıl yapılan bir araştırmada şirketlerin insan kaynakları yöneticilerinin işe alım sürecinde teknolojiyi ne ölçüde kullandıkları sorulmaktadır. Her yıl yaklaşık olarak 500 insan kaynakları personelinin cevapladığı anketin cevaplarına göre bu süreçteki teknoloji kullanımında 2012'de %29 seviyesinde iken 2020'de %68 seviyesine gelmiştir. Şirketlerin %35'i çevrimiçi testler

uygulamakta, %43'ü başvuru takip sistemi kullanmaktadır (CIPD, 2020). Bu tür çevrimçi kaynak verileri saklanarak bunların kestirimci analitik çalışmalarda kullanılması da mümkündür.

Bu alanda veri bilimi ve semantik web teknolojilerinden faydalanılarak geliştirilen projelerden biri de aday özgeçmişlerinin *"web crawler"* yardımı ile derlenmesi, derlenmiş yapısal olmayan verinin yapısal formata getirilmesi, yapısal formattaki veriden çeşitli veri bilimi algoritmaları ile açık pozisyona en uygun adayların seçilmesi şeklindedir. Bu çalışmada Regresyon, K-en yakın komşu, Rastgele orman, Naïve Bayes, Destek vektör makineleri gibi çeşitli yöntemler kullanılarak yapısal hale getirilen veriden adayın eğitim bilgileri, geçmiş iş tecrübesi ve yetenekleri ayrıştırılmaktadır (Necula, S. C. & Strîmbei, C., 2019).

İşe alım sürecine giren adayın değerlendirmesi yapılırken şirket kültürüne uygunluğu, yeni rolünde öğrenme ve gelişme potansiyeli, geçmiş tecrübesi baz alındığında bağımsız şekilde çalışabilme yetkinliği, yeni rol için yetkinliklerinin yeterliliği değerlendilirken de şirketin geçmiş verisinden yola çıkarak bu kişileri değerlendirmesi faydalı olmaktadır. Bunu gerçekleştirebilmek için şirketin mevcut çalışanlarından başarılı ve başarısız gördükleri çalışanları belirli ölçütlere göre değerlendirmeleri ve yeni gelen adayların da benzer testte değerlendirilip sınıflandırılması sağlanabilir. Böylece bir adayın şirkette işe başladığı durumdaki başarı düzeyi tahmin edilebilir (Walford-Wright, G., Scott-Jackson, W., 2018).

3.6.4. İş Başlangıcı, Kültürel Uyum ve Bağlılık Analitiği

Çalışanın işe başlangıcı sonrası uyum sağlaması ve kendinden beklenen performans düzeyine ulaşması zaman alan bir süreçtir ve bu süreç 6 ay ile 1 sene arasında sürmektedir. Bu süreç içerisinde çalışanın karşılaştığı zorluklar, bu zorlukları çözüm yöntemleri, aldığı destekler, işe başlangıç sürecinde aldığı eğitimler, işe devamlılığı, dahil olunan projeler gibi önemli işe başlangıç performans metrikleri tutularak bunların her çalışan için doldurulması sağlanabilir. Sonrasında, saklanan bu veriden çalışanın gelecekteki belirli bir süre içerisinde başarılı veya başarısız olacağı kestirimci istatistiksel modellerle tahminlenebilir (Isson & Harriott, 2016).

Shah, Irani ve Sharif (2017), çalışan davranışları ve tavırlarından Yapısal Eşitlik Modellemesi (SEM) yöntemi ile organizasyonel değişime hazır olma durumunu tespit etmeye çalışmışlardır.

3.6.5. Çalışanın İşten Ayrılması ve Çalışan Muhafazası Analitiği

İnsan kaynakları alanındaki analitik çalışmaların en önemli uygulama alanlarından biri de çalışanı elde tutma amacı taşıyan uygulamalardır. Çalışanların hangi sebeplerle ayrıldığını doğru tespit

etmek ve kalması gerekli olan çalışanları elde tutmak için alınması gereken aksiyonların dizayn edilmesi bu kapsama girmektedir. Aksi takdirde, ayrılması istenmeyen kişilerin kaybedilmesi şirkete yüksek maliyetler yaratabilir (Reilly, 1996).

Şirketlerde işten ayrılmalar aşağıda sıralandığı gibi çeşitli nedenlerle olabilir:

- Emeklilik
- Doğal kayıplar
    - İstifa
    - Hastalık veya yaralanma
    - Ölüm
    - Rıza dışı kayıplar
- Role duyulan ihtiyacın ortadan kalkması
    - Anlaşma karşılığında ilişik kesme
    - Erken emeklilik
    - İşten çıkarma

Bu kayıplardan emeklilik gibi bazıları önceden tahmin edilebildiği için gerekli aksiyonlar alınabilir ve şirketin ilgili kişinin yerini doldurmaya önceden hazırlıklı olması beklenir. Bunun yanında ölüm veya hastalık gibi nedenler ise önlenemeyen durumlardır. Bu kayıplar arasında şirketin elinde olan ve etki edebileceği en önemli madde istifalardır. İstifalar, hem kaybedilmemesi gereken bir çalışanın kaybından dolayı hem de yerini doldurma işlemlerinden dolayı şirketlere ek maliyetler doğurmaktadır (Reilly, 1996).

Kishore, vd. (2012), yazdıkları makalede veri analitiğinin yetenekli çalışanların muhafazasına odaklandığını ve güvenilir veri ile veriye dayalı kararın bu alanı yeniden inşa edeceğini belirtmişlerdir. Makalede çalışan memnuniyetini sağlayıp onların işten ayrılmasını ve dolayısıyla işgücü devir oranını azaltmak amacıyla çalışan görüşmeleri yapılması önerilmektedir. Ayrıca çalışanların rollerindeki beklentilerinin ve kariyer yollarının açıkça ifade edilmesinin de bu amaca hizmet ettiği belirtilmektedir.

Tezin ilerleyen bölümlerinde bu konuya daha detaylı değinilecektir.

3.6.6. Çalışan Performans Yönetimi ve Ömürboyu Değer Hesabı Analitiği

İnsan kaynakları alanında mevcut çalışanları performans yönetimi hem eksik veri, subjektif ve kalitatif metrikler nedeniyle hem de performansı ölçmenin zorluğu nedeniyle en problemli alanlardan biridir. Geçmişte TOPSIS, parametrik olmayan yöntemler, bulanık sinir ağları ve çeşitli sıralama

yöntemleri kullanılarak bu alanda çalışmalar yapılmıştır (Kalhori, M. & Kargar, M., J. (2017). Yapılan çalışmalar ile veriye dayalı olarak performans değerlendirmesi yapılan çalışanların performans gelişim alanları tespit edilerek gerekli yönlendirmeler yapılmalıdır. Bu yönlendirmeler sonucunda da performans izleme aşamasına geçilmelidir. Jantan, Hamdan ve Othman (2011) ise çalışanın geçmiş bilgilerinden gelecekteki performansının tahmin edilerek ona göre bir performans yönetimi yapılmasına yönelik bir konferans raporu hazırlamışlardır.

Yapılan analitik performans yönetimi çalışmaları ile insan kaynakları yönetiminde büyük bir dönüşüm olması beklenmektedir ve bunun da şirketin çıktılarına önemli ölçüde etki etmesi beklenmektedir. Geleneksel performans yönetimi ile analitik performans yönetimi arasındaki farklar bir tablo ile aşağıdaki şekide özetlenebilir. Analitik yöntemler ile performans yönetimi ile işe alınan yeni çalışanlarda %33 oranında daha yüksek başarı oranı yakalandığı görülmüştür (Isson & Harriott, 2016).

Tablo 9. Geleneksel ve Analitik İnsan Kaynakları Yönetimi

	Geleneksel Yöntemler	Analitik Yöntemler
Terfi	• Kalitatif • Kıdeme dayalı • Şeffaf olmayan	• Kantitatif • Performansa dayalı • Şeffaf • Analitiğe dayalı kariyer yolu
Ücretlendirme	• Yılda bir gözden geçirilen • Rating notuna göre karar • Kısıtlı değerlendirme	• Sürekli değerlendirilen • Etkiye dayalı karar • Daha geniş değerlendirme
Çalışan geri bildirimleri	• Yıllık geri bildirim • Reaktif • Üst düzey yorumlar	• Aylık geri bildirim • Proaktif ve karşılıklı • Detaylı ve spesifik

Saradhi, V., V. ve Palshikar, G., K. (2011) tarafından önerilen bir yaklaşıma göre de tıpkı müşteri ömür boyu değeri hesaplandığı gibi çalışan için de ömür boyu değer hesabı yapılabilir ve bu performans metriğine göre kaybedilmemesi elzem olan çalışanlar veriye bağlı olarak belirlenebilir. Bu çalışanların şirkete değeri çok daha fazla olacağından diğer çalışanlara göre daha özel bir ilgi gösterilmesi gerekir.

3.6.7. Çalışan Sağlığı ve Güvenliği Analitiği

Çalışanın sağlığı ve güvenliği hem ürün veya hizmetin daha başarılı olması ile birlikte şirketin finansal çıktılarına etki etmekte hem de çalışanın şirkette daha uzun süreli çalışması konusunda etkin

rol oynamaktadır. Rosen ve Spaulding'e göre (2009) insanların sağlık harcamalarının %75'inin doğru davranış ile önlenebileceği belirtilmektedir. Deloitte tarafından 2019 yılında yayınlanan bir makalede de çalışan sağlığı alanında kestirimci analitik yöntemlerin uygulanması ile birlikte sağlık sorunlarının önceden tahmin edilip önlenebileceği söylenmiştir. Bunun yanında insanların yaptığı bazı işlerin makineler tarafından yapılacak şekilde otomatize edilmesi de önemli rol oynamaktadır.

İşyerinde kestirimci analitik yöntemler ile çalışan sağlığı alanında yapılacak çalışmalar aşağıdaki spesifik alanlarda olabilir (Isson & Harriott, 2016):

- Çalışanların çalıştıkları sektör, iş sahası ve iş nedeniyle yaralanma ihtimalinin tahmini ve önlenmesi için aksiyonların alınması
- Çalışanların hastalık izni alma ihtimalinin tahmini ve buna göre işlerin yedeklenmesi ve diğer çalışanlara aktarımının sağlanması
- Hangi araç, makine ve teçhizatın yakın zamanda bozulacağının tahmini ve buna göre bakım planlaması yapılması
- Hangi dönemde sigortaya yapılan geri ödeme bildirimlerinin en yüksek düzeye ulaşacağının tahmini ve buna göre maliyet yönetimi

4. ÇALIŞAN İŞTEN AYRILMASI

Bu tez çalışması insan kaynakları analitik uygulama alanlarından çalışanın işten ayrılması ve çalışan muhafazası alanlarına odaklanmaktadır ve bu alandaki çalışmalar daha detaylı şekilde incelenmiştir. Bu bölümde çalışanın işten ayrılması konsepti, hesaplamaların nasıl yapıldığı, geçmişte yapılan araştırmalarda çalışanın işten ayrılmasının en önemli sebeplerinin neler olduğu anlatılmaktadır. Bir sonraki bölümde ise bu alan özelinde yapılan bir uygulama anlatılmıştır.

4.1. Konseptler

İş yaratımı ve iş kaybının ölçümü ve yorumlanması için tanımlar ve varsayımların net olması gerekmektedir. Terminolojide "iş"in çalışan tarafından doldurulan bir pozisyon olduğu düşünüldüğünde iş yaratımı ve iş kaybı şu şekilde tanımlanabilir (Davis, S. J. & Haltiwanger, J., 1999):

Tanım 1: t anında yaratılan brüt iş t-1 ile t zamanları arasında tüm iş birimlerinde genişletilen veya yeni açılan işlerin toplamıdır.

Tanım 2: t anında kaybedilen brüt iş t-1 ile t zamanları arasında tüm iş birimlerinde daraltılan veya kapatılan işlerin toplamıdır.

Bu iki tanım göz önüne alındığında net işgücü değişimi, basitçe iki tanımın farkı alınarak hesaplanabilir.

Tanım 3: t anında yer değiştiren brüt iş t-1 ile t zamanları arasında tüm iş birimlerinde yaratılan veya kaybedilen işlerin toplamıdır. Yani, Tanım 1 ve 2'nin toplamıdır.

4.2. Hesaplamalar

Bahsi geçen konseptleri formülize etmek anlamayı ve aralarındaki bağlantıyı kurmayı kolaylaştıracaktır. EMP_{est}, s sektöründeki e işvereninde t anındaki çalışan sayısını gösteriyor olsun. Diğer yandan, S_t t anındaki işverenlerin kümesi, S_t^+ t-1 ile t anları arasında genişleyen veya yeni açılan işverenlerin kümesi, S_t^- ise t-1 ile t anları arasında daralmaya giden veya kapatılan işverenlerin kümesi olsun (Davis, S. J. & Haltiwanger, J., 1999).

s sektöründe t anındaki brüt iş yaratımı (Tanım 1'e göre);

$$C_{st} = \sum_{e \,\epsilon\, S^+} \Delta EMP_{est}$$

s sektöründe t anındaki brüt iş kaybı (Tanım 2'ye göre);

$$D_{st} = \sum_{e \,\epsilon\, S^-} |\Delta EMP_{est}|$$

Sektörel net iş yer değiştirmesi ise şu şekilde hesaplanır (Tanım 3'e göre).

$$R_{st} = \sum_{e \,\epsilon\, S} |\Delta EMP_{est}| = C_{st} + D_{st}$$

İşten ayrılmaları formülize etmek için bazı ek hesaplamalar da yapılması gerekmektedir. t ve t-1 anlarındaki toplam işgücünün ortalaması alınarak, t anındaki iş kayıpları ile oranına bakıldığında işten ayrılmaların oranı bulunabilir. Aşağıdaki formüllerde Z_{est}, s sektöründeki e işvereninde t ve t-1 anlarındaki işgücü büyüklüklerinin ortalamasını ifade etmektedir. Z_{st} ise işverenden bağımsız olarak tüm sektör için bir hesaplamadır. Bunlardan hareketle, işen ayrılmalar, d_{st}, oran olarak hesaplanabilir (Davis, S. J. & Haltiwanger, J., 1999).

$$Z_{est} = 0.5\,(EMP_{est} + EMP_{es(t-1)})$$

$$Z_{st} = \sum_{e \,\epsilon\, S} Z_{est}$$

$$d_{st} = \frac{D_{st}}{Z_{st}}$$

4.3. Çalışan İşten Ayrılmasının En Önemli Sebepleri

İşten ayrılmaların nedenleri konusunda birçok çalışma yapılmıştır ve bunlar verinin ait olduğu ülke, verinin seviyesi (ülke, şirket, çalışan vb.), uygulanan yöntem ve sonuçları bakımından farklılaşmaktadır. Fitz-enz, J. (2012), yaptığı bir çalışmada çalışanların işten ayrılma sebeplerini şu şekilde sıralamıştır:

- Adil olmayan maaş artışları
- Adil olmayan performans değerlendirme süreci
- Adil olmayan terfi uygulamaları
- Adil olmayan kural, politika ve süreçler
- Aşırı iş yükü
- Bürokrasi
- Cinsel taciz

- Çalışanı düşünen liderlik eksikliği
- Çalışanlar arasında adam kayırıcılık
- Daha yüksek maaşlı bir iş bulunması
- Direkt amir ile ile ilgili sorunlar
- Düşük performansa aksiyon alınmaması
- Düşüncelerin hoş karşılanmaması
- Emeklilik
- Eski ve yetersiz ekipman
- Esnek olmayan çalışma saatleri
- Eş durumu nedeniyle taşınma
- Etik veya dürüst olmayan yöneticiler
- Fiziksel çalışma ortamındaki ses, kir, sıcaklık gibi nedenler
- Irk, cinsiyet, din ve diğer konulardaki ayrımcılık
- İstenmeyen yer değişimi
- İstenmeyen taşınma
- İş arkadaşlarıyla ilgili sorunlar
- İş güvenliği
- İş kaybı
- İş sorumlulukları
- İş-özel yaşam dengesinin eksikliği
- İşe mikro seviyede müdahil olunması
- İşe ulaşmanın zorluğu ve işyeri mesafesi
- İşi kaybetme korkusu
- İşi yapabilecek yetki eksikliği
- İşi yapabilme yetkinliğinin olmaması
- İşi yapmak için kaynak yetersizliği
- İşin coğrafi lokasyonu
- İşin kendini geliştirmeye açık olmaması
- İşin kendisi
- İşteki istenmeyen değişimler
- İzin politikası

- Kariyer yolu değişimi
- Kariyer yolunun tıkanması
- Kaygıları dile getirmeye izin verilmemesi
- Kazanç potansiyelinin düşük olması
- Kendi işini kurma
- Negatif çalışma ortamı
- Okula geri dönüş
- Organizasyon kültürü
- Organizasyonun stabil olmaması
- Organizasyon politikaları
- Öğrenme fırsatı olmaması
- Performansa dayalı ücretlendirme yapılmaması
- Performansa dayalı takdir edilmeme
- Performans geri bildirimi eksikliği
- Prim sistemi olmaması
- Rekabetçi olmayan ücret politikası
- Sağlık kaygıları
- Stres
- Sürekli değişim olması
- Şirketin geleceği hakkındaki kaygılar
- Takım çalışması olmaması
- Ücret artışlarının zamanında olmaması
- Üst yönetime güven duyulmaması
- Yan hakların yetersizliği
- Yetersiz veya uygun olmayan eğitimler
- Yetkilendirme eksikliği
- Yükselme şansı olmaması
- Zayıf değerlendirme
- Zayıf iletişim

Bu nedenlerden bir kısmı üzerinde daha detaylı çalışmalar yapılmış ve spesifik olarak kitaplar ve makalelere konu olmuştur. Bu bölümün devamında bu alanda yazılan makale ve kitaplara atıf yapılarak işten ayrılma nedenlerinin detayları üzerinde durulacaktır.

4.3.1. Demografik Özellikler

Çalışmaların bir kısmı işten ayrılmaları açıklayabilecek en önemli değişkenlerin yaş, cinsiyet gibi demografik özellikler olduğunu öne sürmüştür. Urwin, P. ve Parry, E. (2016) 20 yıllık bir zaman diliminde Birleşik Krallık'taki şirketler için kıdem ve işten ayrılmalar arasındaki ilişki üzerine çalışmıştır. Demografik bilgiler, yaş, cinsiyet, ekonomik resesyon gibi bilgiler de işten ayrılmanın açıklanabilmesi için kullanılmıştır. Davis ve Haltiwanger (2014) de Amerika'da yüksek işten ayrılma oranlarının genç ve daha düşük eğitimli kesimde daha yüksek olduğu ve aynı yaş ve eğitim seviyesindeki çalışanlarda cinsiyetin belirleyici bir rol oynamadığı sonucuna ulaşmışlardır.

OECD'nin 2009 İstihdam raporuna göre uluslararası iş piyasaları incelendiğinde İngiltere ve Amerika'daki iş değişiklikleri %25-45 arasında iken bu oran Avrupa'da çok daha düşüktür. Ayrıca aynı raporda, genç çalışanların, kadınların ve daha az kalifiye iş yapanların daha mobil oldukları bilgisi de verilmektedir. OECD'nin 2010 raporunda ise iş güvencesi ve kadrolu çalışmanın işten ayrılmalar üzerinde ters etkisi olduğu bilgisi rakamlar ile desteklenmektedir.

Çalışanın işten ayrılmasının tahmini konusunda ileri analitik yöntemlerin uygulanması da son yıllarda ağırlık kazanmıştır. Pérez-Campdesuñer, vd. (2018) doğrudan işten ayrılmaları tahmin edecek bir model geliştirmek amacıyla sinir ağı algoritmalarını kullanmışlardır. Bu çalışmanın sonucunda gelir, eğitim düzeyi, yaş gibi faktörler en anlamlı faktörler olurken sektör, sektördeki tecrübe, şirkette çalışılan yıl veya mevcut pozisyonda çalışılan süre gibi faktörler anlamlı çıkmamıştır.

Saradhi, V., V. & Palshikar, G., K. (2011) tarafından yapılan bir araştırmada Destek vektör makineleri yöntemi ile çalışanların işten ayrılması incelenmiştir ve kurulan kestirimci analitik modelinin en önemli değişkenlerinden ikisinin yaş ve cinsiyet olduğu görülmüştür.

4.3.2. İş Memnuniyeti

Bazı çalışmalar ise iş memnuniyetinin işten ayrılmaların tahmininde daha açıklayıcı olduğunu savunmuştur. Almanya'da yapılan bir çalışmada ise Lévy-Garboua, vd. (2007) iş memnuniyeti ile işten

ayrılmalar arasında bir bağıntı kurmaya çalışmışlardır ve bu amaçla 1985-2003 yılları verisini kullanmışlardır. Clark (2001), İngiltere'de yaptığı araştırmada İngiltere Hanehalkı Panel Verisi'nden hareketle işten ayrılmaları açıklamaya çalışmış ve iş memnuniyetinin en önemli etken olduğu sonucuna varmıştır. İş memnuniyetini etkileyen en önemli kriterler olarak da iş güvenliği ve maaş öne çıkmıştır.

Böckerman ve Ilmakunas (2009) ise yaptıkları araştırmada iş memnuniyeti ve çalışma koşullarının çalışanları bir işte çalışırken iş aramaya itebileceğini ve bunun da işten ayrılma sonucunu doğuracağını öne sürmüştür. Dale-Olsen (2016) de çalışanların işten ayrılması konusunu çalışan motivasyonu, çalışan tercihleri ve çalışanın karşısına çıkan fırsatlar ile ilişkilendirmektedir. Lee, vd. (2018), aynı şekilde iş memnuniyetindeki azalma ve alternatif iş fırsatlarının çalışanın kaybında etkili en önemli parametreler olduğu sonucuna ulaşmıştır. Stamolampros, P. vd. (2019) tarafından yapılan bir çalışmada 297 bin çevrimiçi değerlendirme sonucuna dayanarak çalışanların işten ayrılma nedenleri sorulduğunda da benzer şekilde iş memnuniyetinin en önemli faktörlerden biri olduğu ve bunun da şirket içerisindeki kariyer fırsatlarından etkilendiği sonucu çıkmıştır.

Yang, J. T., Wan, C. S. & Fu, Y. J. (2012) yaptıkları anket çalışmasında Tayvan'da turizm endüstrisinde çalışan yönetici seviyesindeki personelin işten ayrılma sebeplerini araştırmışlardır. Bu araştırma sonucunda anket katılımcılarının %80'inin daha iyi bir kariyer hedefi nedeniyle, %70'inin ise rakip şirketlerde daha tatmin olacağı düşüncesiyle iş değiştirmeyi düşündüğü sonucu çıkmıştır. Afrika'da yapılan benzer bir çalışmada da bir banka çalışanları üzerinde yapılan anket ve bu anket sonuçlarının Yapısal Eşitlik Modellemesi ile analizi sonucunda iş tatmini ve memnuniyetinin en önemli sebep olduğu sonucuna ulaşılmıştır (Sabbagha, M. D. S., vd., 2018). Avutralya ve Yeni Zelanda'lı araştırmacılar Aliyu, O. A. ve Nyadzayo, M. W. tarafından (2018) Malezya'daki çağrı merkezi çalışanları üzerinde yapılan bir çalışmada da benzer sonuçlara ulaşılmış ve iş memnuniyetinin en önemli bir faktör olduğu ortaya çıkarken bunu etkileyen en önemli sebep olarak da çağrı merkezi şirketleri özelinde müşteri ilişki yönetimi (CRM) stratejisinin bunu sağladığı belirtilmiştir. Bu çalışmada da Yapısal Eşitlik Modellemesi kullanılmıştır. 2018 yılında Fletcher, L., Alfes, K. ve Robinson, D. tarafından yapılan ve "Russell's Model" yönteminin kullanıldığı araştırmada da iş memnuniyeti, çalışanın motivasyonu ve stres en önemli faktörler olarak karşımıza çıkmıştır. Bu çalışmada, iş memnuniyetini sağlayan etken olarak da eğitim ve gelişim alanına odaklanılmıştır.

4.3.3. Maddi Sebepler

Tahmin edilebileceği gibi, işten ayrılmaların nedenini maaş seviyesi olarak saptayan çalışmalar da mevcuttur. Delfgaauw (2007), Hollanda'da yaptığı ve ankete dayanan bir araştırmada

kamu sektörü çalışanlarının işten ayrılma nedenlerini incelemiştir. Elde edilen sonuçlar maaş ve iş baskısı nedeniyle işten ayrılan çalışanların sektör değiştirdiği, ancak iş atmosferini beğenmedikleri için işten ayrılanların ise sektör içerisinde kaldığını ortaya çıkarmıştır. Dale-Olsen (2006) yüksek maaş ve daha fazla yan hakların işten ayrılma oranını düşürdüğü sonucuna varmıştır.

Alao ve Adeyemo (2013) karar ağaçları kullanarak Nijerya'daki yüksek öğretim kurumlarındaki çalışanların 28 yıllık verisi üzerinde işten ayrılma tahmini yapmıştır. Araştırmacılar, uzun yıllardır çalışan ve maaş seviyelerindeki artış düşük olan çalışanların şirketten ayrılma ihtimalinin daha yüksek olduğu sonucuna ulaşmışlardır.

4.3.4. Çalışma Koşulları

Ayrıca, şirket ve çalışma koşulları ile ilgili nedenlerin işten ayrılmada açıklayıcı olduğunu savunan çalışmalar da mevcuttur. Bellon (2016), Fransa'da yaptığı bir çalışmada işten ayrılmaları şirket büyümesi ile ilişkilendirmiş ve büyüyen şirketlerin zamanla ilgili role daha uygun adaylar tercih etmesi ile birlikte işe alım ve işten ayrılmaların arttığını gözlemlemiştir. Burgess, Lane ve Stevens (2000), yaptığı çalışmada işten ayrılmaların sektöre ve şirketin yaşına bağlı olduğu sonucuna ulaşmıştır. Yine aynı araştırmacılar 2001 yılında yaptıkları başka bir çalışmada da şirket seviyesinde incelemelerde bulunmuş ve bu tezlerini doğrulayıcı sonuçlara ulaşmışlardır.

Katsikea, E., vd. (2015), satış yöneticileri arasında yaptıkları bir araştırmada satış yöneticilerinin işten ayrılmasından en önemli tetikleyicinin iş stresindeki artışa bağlı olarak memnuniyetin düşmesi olduğu sonucuna ulaşmışlardır. Lee, vd. (2018), makalesinde Gallup tarafından hazırlanan bir raporda çalışanların %51'inin esnek çalışma saatleri için iş değiştirmeye sıcak bakacağını bildirmiştir.

Cottini, Kato ve Westergaard-Nielsen (2011), Danimarka'da çalışan-işyeri eşleştirilmiş verisi üzerinde logit modelleme ile yaptıkları çalışmada kötü çalışma koşullarının işten ayrılmada en önemli etken olduğunu bulmuşlardır. Fiziksel tehlikeye maruz kalınan işyerlerinde işten ayrılma oranında %3, gece vardiyası ile çalışılan yerlerde işten ayrılma oranında %11 ve destek olmayan bir patron ile çalışılan durumlarda da işten ayrılma oranında %6 artış görmüşlerdir.

4.3.5. Şirket – Çalışan Uyumsuzluğu

Presbitero, A., vd. (2016), çalışanlar ile şirket değerlerinin birbirine uyumunun önemine dikkat çekecek bir çalışma yapmış ve Filipinler'deki taşeron firmalarda 258 çalışan ile yaptıkları

araştırma sonucunda çalışan işe şirket değerlerinin uyumu ile çalışanların işten ayrılması arasında önemli derece bir ilişki yakalamışlardır.

4.3.6. Uygun Olmayan İşe Alım Süreci

Hendrick, R. Z. ve Raspiller, E. E. tarafından yapılan bir çalışmada (2011), çalışanların işe alınırken doğru değerlendirilmemesinin çalışanların 12 ay içerisinde işten ayrılmasında ayırt edici faktör olduğu belirtilmiştir. Bu araştırmaya göre, uygun bir ön değerlendirme sürecine girip başarıyla geçen çalışanların işyerinde çalışma süresi diğerlerine göre daha uzundur.

Çalışanların işten ayrılma sebeplerini araştırmak dışında ileri analitik yöntemleri birbirleriyle karşılaştıran makaleler de yayımlanmıştır. Punnoose ve Ajit (2016) tarafından yapılan çalışma da benzer şekilde işten ayrılma tahminini makine öğrenmesi yöntemleriyle yapmaya çalışmış ve çeşitli yöntemleri birbiri ile karşılaştırmıştır. Karşılaştırma sonunda Extreme Gradient Boosting metodunun eldeki veri için en iyi performansı sağladığı görülmüştür. Bhuva ve Srivastava (2018) ile Salunkhe (2018) yine insan kaynağı kaybını tahminleyen çeşitli modeller kurarak performans karşılaştırması yapmıştır. Türkiye'de de benzer alanda yapılan bir çalışma bulunmaktadır. Bu çalışma Kısaoğlu (2014) tarafından yapılmış olup bu çalışmada çalışanların şirket içi bilgileri yerine profesyonel sosyal medya aracı üzerindeki işe giriş-çıkış bilgileri ile iş bilgileri kullanılmıştır.

Aşağıdaki tabloda çalışan işten ayrılmasını tahmin etmek amacıyla yazılan makaleler ve bunların yöntem ve yıl bazında sınıflandırması yer almaktadır.

Tablo 10. Literatür Sınıflandırması

Yazarlar	Yıl	Problem	Yöntem
Burgess, S., Lane, J. & Stevens, D.	2001	Çalışan işten ayrılması	Empirik Analiz
Clark	2001	Çalışan işten ayrılması	Empirik Analiz
Dale-Olsen, A.	2006	Çalışan işten ayrılması	Empirik Analiz
Delfgaauw	2007	Çalışan işten ayrılması	Empirik Analiz
Lévy-Garboua, vd.	2007	Çalışan işten ayrılması	Standart Ekonomi Teorisi
Böckerman & Ilmakunnas	2009	Çalışan işten ayrılması	Probit model

Hendrick, R. Z. ve Raspiller, E. E.	2011	Çalışan işten ayrılması	Empirik Analiz
Jantan, Hamdan & Othman	2011	Çalışan performans yönetimi	Karar ağaçları
Cottini, Kato & Westergaard-Nielsen	2011	Çalışan işten ayrılması	Lojistik regresyon
Saradhi, V., V. & Palshikar, G., K	2011	Çalışan işten ayrılması	Destek vektör makineleri
Kishore, K., Majumdar, M, & Kiran	2012	Çalışan işten ayrılması	İnceleme
Yang, J.T., vd.	2012	Çalışan işten ayrılması	Empirik Analiz
Ünal, Ö. F.	2012	Çalışan performans yönetimi	AHP
Alao & Adeyemo	2013	Çalışan işten ayrılması	Karar ağaçları
Davis & Haltiwanger	2014	Çalışan işten ayrılması	Empirik Analiz
Katsieka, E., Theodosiou, M. & Morgan, R., E.	2015	Çalışan işten ayrılması	Empirik Analiz
Urwin & Parry	2016	Çalışan işten ayrılması	Empirik Analiz
Bellon	2016	Çalışan işten ayrılması	Lojistik regresyon
Punnoose, R., Ajit, P	2016	Çalışan işten ayrılması	Makine öğrenmesi
Kalhori & Kargar	2017	Çalışan performans yönetimi	TOPSIS, FAHP
Shah, N., Irani, Z. & Sharif A	2017	Organizasyonel değişime hazırlık	Yapısal eşitlik modellemesi
Presbitero, A., Roxas, B., & Chadee, D.	2017	Çalışan işten ayrılması	Yapısal eşitlik modellemesi
Yiğit, İ. O., Shourabizadeh, H.	2017	Çalışan işten ayrılması	Makine öğrenmesi
Walford-Wright, G., Scott-Jackson, W.	2018	Yetenekli çalışan işe alımı	İnceleme
Lee, T. W., vd.	2018	Çalışan işten ayrılması	İnceleme
Sabbagha, M. D. S., vd.	2018	Çalışan işten ayrılması	Yapısal eşitlik modellemesi

Aliyu, O. A. & Nyadzayo, M. W.	2018	Çalışan işten ayrılması	Yapısal eşitlik modellemesi
Fletcher, L., Alfes, K. & Robinson, D.	2018	Çalışan işten ayrılması	Russell's model
Pérez-Campdesuñer, vd.	2018	Çalışan işten ayrılması	Makine öğrenmesi
Bhuva & Srivastava	2018	Çalışan işten ayrılması	Makine öğrenmesi
Salunkhe	2018	Çalışan işten ayrılması	Makine öğrenmesi
Alamsyah, A., Salma, N.	2018	Çalışan işten ayrılması	Makine öğrenmesi
Attri, T.	2018	Çalışan işten ayrılması	Makine öğrenmesi
Noack, B.	2019	İşe alım	İnceleme
Gürkan, S., Dazlak, B. D.	2019	İşe alım	AHP
Stamolampros, P., vd.	2019	Çalışan işten ayrılması	Lojistik regresyon
Necula, S. C. & Strîmbei, C.	2019	Yetenekli çalışan işe alımı	Makine öğrenmesi

5. ÇALIŞAN KAYBI ANALİTİĞİ VE OPTİMİZASYONU UYGULAMASI

Tez kapsamında yapılacak uygulama insan kaynakları analitiğinin son yıllarda popüler olan konularından biri olan çalışanın işten ayrılmasını önceden tahmin edebilme üzerinedir. Bu amaçla, akademik olarak kullanılmasında sakınca olmayan anonim bir veri kümesi kullanılarak çalışma yapılmıştır. Bu bölümdeki diğer alt başlıklarda çalışmanın detayına yer verilmiştir.

5.1. Veri Kümesi

Çalışmada kullanılması önerilen veri kümesi bir teknoloji şirketinin akademik ve deneysel çalışmalar için kullanılmak üzere paylaştığı ve içerisinde aşağıdaki alanları da içeren bir veri kümesidir. Görüleceği gibi veri içerisinde çalışana ait hem demografik bilgiler hem de iş ile ilgili bilgiler mevcuttur. Çalışma sırasında bu alanlara ek olarak başka değişkenler de türetilerek modellemede kullanılacaktır.

Tablo 11. Kullanılacak Veri Kümesi

Girdi Değişkenler		
Yaş	Cinsiyet	Medeni Durum
Eğitim	Eğitim Alanı	
Departman	İş Rolü	İş Seviyesi
Şirkette çalışılan süre	Mevcut rolde çalışılan süre	Toplam tecrübe
Mevut yöneticiyle çalışılan süre	Son terfiden geçen süre	Çalışılan firma sayısı
Yüzde maaş artışı	Performans notu	Eğitim sayısı
İş-hayat dengesi	İşe dahiliyet	Standart saatler
Fazla mesai durumu	Hisse seviyesi	Ev-İş Mesafesi
İş seyahati		
İş memnuniyeti	Ortam memnuniyeti	İlişki memnuniyeti
Saatlik ücret	Günlük ücret	Aylık ücret
Aylık Gelir		
Hedef Değişken		

İşten Ayrılma (0/1)

5.2. Veri Hazırlığı ve Ön İşlemleri

Çalışmaya başlamadan önce verideki tüm alanlar tezin diline de uygun halde olması için Türkçe olarak çevrilmiştir. Sonrasında, daha anlaşılır olması açısından verinin demografik veya işle ilgili olmasına göre değişkenler kendi içinde sıralanmıştır.

Veri, 1470 satır ve 35 sütundan oluşmaktadır. Yani, 1470 çalışanın 34 özelliği ile 1 hedef değişken tahmin edilmeye çalışılacaktır.

5.3. Veri Analizi

Veri alanlarının türleri incelenmiş ve hem sayısal hem de kategorik verilerin yer aldığı gözlemlenmiştir. Aşağıdaki tabloda her değişkenin alan türü ve doluluk oranı gösterilmiştir. Tabloda, her değişken için 1470 kaydın hepsinin dolu olduğu görülmektedir; yani boş değer bulunmamaktadır. Ayrıca, her değişkenin kaç farklı değer aldığı da tabloya eklenmiştir. Bu değerlerin beklentiler dahilinde olduğu söylenebilir. Ancak sadece tek bir değer gelen değişkenlerin modelleme anlamında bir katkısı olmayacağından bu değişkenler sonraki adımlarda tablodan silinecektir.

Tablo 12. Veri Tipleri

Değişken	Değişken Tipi	Tekil Değer
Calisan_No	int64	1470
Yas	int64	43
Yas_18_Ustu	object	1
Cinsiyet	object	2
Medeni_Durum	object	3
Egitim_Duzeyi	int64	5
Egitim_Alani	object	6
Calisan_Sayisi	int64	1
Departman	object	3
Is_Rolu	object	9
Is_Seviyesi	int64	5
Sirkette_Calisilan_Sure	int64	37
Mevcut_Rolde_Calisilan_Sure	int64	19

Toplam_Tecrube	int64	40
Mevcut_Yoneticiyle_Calisilan_Sure	int64	18
Son_Terfiden_Gecen_Sure	int64	16
Calisilan_Firma_Sayisi	int64	10
Yuzde_Maas_Artisi	int64	15
Performans_Notu	int64	2
Egitim_Sayisi	int64	7
Is_Hayat_Dengesi	int64	4
Ise_Dahiliyet	int64	4
Standart_Saatler	int64	1
Fazla_Mesai	object	2
Hisse_Seviyesi	int64	4
Ev_Is_Mesafesi	int64	29
Is_Seyahati	object	3
Is_Memnuniyeti	int64	4
Ortam_Memnuniyeti	int64	4
Iliski_Memnuniyeti	int64	4
Saatlik_Ucret	int64	71
Gunluk_Ucret	int64	886
Aylik_Ucret	int64	1427
Aylik_Gelir	int64	1349
Isten_Ayrilma	object	2

Sayısal olan değişkenlerin özet istatistikleri incelenmiştir (adet, ortalama, standart sapma, minimum, maksimum ve yüzdelik değerler) ve aşağıda tablo olarak eklenmiştir. Sonrasında histogram grafikleri çizilerek de analiz edilmiştir.

Özet istatistik tablosu ve histogram grafikleri birlikte incelendiğinde bulgular şunlar olmuştur:

- Çalışanlar 18-60 yaş arasında dağılmaktadır ve dağılım sağa çarpıktır; ortalama yaş 37 yaş civarındadır.
- Eğitim seviyesi sayısal değerlerle belirlenmiştir ve kitlenin büyük çoğunluğu 3 ve 4 gruplarında toplanmıştır.
- Çalışan sayısı alanı her kayıt için aynı değerden oluşmakta olduğundan ayırt edici bir bilgi değildir ve veri içerisinde tutulmamalıdır.
- Eğitim seviyesi sayısal değerlerle belirlenmiştir ve ortalaması 2.06 olacak şekilde kitlenin büyük çoğunluğu 1 ve 2 gruplarında toplanmıştır.
- Şirkette çalışılan süre ortalamada 7 yıl olmakla birlikte kitlenin yarısı 5 yıldan daha kısa süredir firmada çalışmaktadır. Bununla birlikte 40 yıldır şirkette çalışan çalışanlar da bulunmaktadır.
- Çalışanlar ortalamada 4.23 yıldır mevcut rollerinde çalışmaya devam etmektedirler. Bununla birlikte dağılım sağa çarpıktır ve çalışanların genelde 3-4 yılda bir rollerinin değiştiği yorumu yapılabilir. Maksimum değer olarak 18 yıldır aynı rolde çalışan çalışanlar da bulunmaktadır.
- Personelin toplam tecrübesi 11.28 yıl olup yaş dağılımı ile birlikte değerlendirildiğinde çalışanların yarısının tecrübesinin 10 yıl altında olması normal karşılanmaktadır.
- Çalışanlar ortalamada 4.12 yıldır mevcut yöneticileri ile çalışmaya devam etmektedirler. Bununla birlikte dağılım sağa çarpıktır ve bu değişken de mevcut rolde çalışma süresiyle birlikte değerlendirilmelidir. Çalışanların rolleri değiştikçe yöneticileri de değişmektedir.
- Çalışanların son terfilerinden itibaren geçen süre ortalamada 2.19 yıldır. Ayrıca, kitlenin yarısı için bu değer 1'den küçük ve %75'inde 3'ten küçüktür. Yani, şirket çalışanların büyük çoğunluğuna kısa sürede terfi vermektedir.
- Çalışanların şu ana kadar kaç farklı firmada çalıştığına bakıldığında ortalamada 2.69 farklı firmada çalıştıkları görülmektedir. Grafik yine sağa çarpık olmakla birlikte maksimum değer olarak 9 farklı firmada çalışmış olan personel de bulunmaktadır.
- Çalışanların yüzde maaş artışları ortalamada %15.21 olarak gerçeklemiştir. Minimum değer %11, maksimum değer %25 olup kitlenin en düşük ve yüksek %25'lik kitleleri haricindeki kısmı %12-%18 bandında zam almıştır. Dolayısıyla en yüksek zam alan kısım dışındaki kitlede zammın çok ayrışmadığı yorumu yapılabilir.

- Çalışanların performans notları sadece 3 ve 4 değerlerini almıştır ve ortalaması 3.15'dir. Bu bilgi ışığında olağan dışı şekilde iyi performans gösteren çalışanların kısıtlı sayıda olduğu görülmektedir.
- Çalışanlara şirketin verdiği eğitimlerin ortalama sayısı 2.80 olmakla birlikte maksimum 6 eğitim almış çalışan bulunmaktadır. Kitlenin %75'i ise en fazla 3 eğitim alabilmiştir.
- İş ve hayat dengesi 1-4 arasında değerlendirilmiştir ve en fazla 3 değeri görülmektedir. Ortalama değer ise 2.76'dır.
- İşe dahiliyet verisi de iş hayat dengesi değişkeni ile neredeyse aynı dağılıma sahiptir.
- Standart saatler değişkeni tüm veride aynı değerde gelmiş olduğundan modellemede ayırt edici olmayacaktır. Bu sebeple, veriden temizlenecektir.
- Çalışanların hisse seviyeleri 0-3 arasında dağılmaktadır ve kitlenin büyük kısmı 0 ve 1 değerlerinde yoğunlaşmıştır.
- Ev-iş mesafesi 1 ile 29 arasında değer almış olup ortalama değer 9.19'dur. Bu grafikte de sağa çarpıklık vardır ve 10 altı değerler yoğunlaşmıştır.
- İş memnuniyeti değişkeni; iş-hayat dengesi ve işe dahiliyet değişkenleri ile çok benzer bir dağılıma sahiptir. Aralarındaki tek fark ise bu değişkenin standart sapmasının diğer ikisine göre nispeten daha yüksek olmasıdır.
- Ortam memnuniyeti değişkeni ile iş memnuniyeti değişkenleri birbirine çok benzemektedir. Çalışanların bu sorulara aynı cevabı vermiş olma ihtimali yüksektir.
- İlişki memnuniyeti değişkeni de ortam ve iş memnuniyeti değişkenleri ile benzerdir.
- Saatlik ücret 30 ve 100 arasında değişen değerler almaktadır ve sola çarpık bu veride ortama değer 65.89 olarak gerçekleşmiştir.
- Saatlik ücret gibi günlük ve aylık ücretler de tekdüze dağılım özelliği göstermektedir.
- Aylık gelirin ortalamada 6503 dolar olduğu, minimum ve maksimum gelirlerin ise sırasıyla 4708 ve 19999 dolar oldukları görülmektedir.

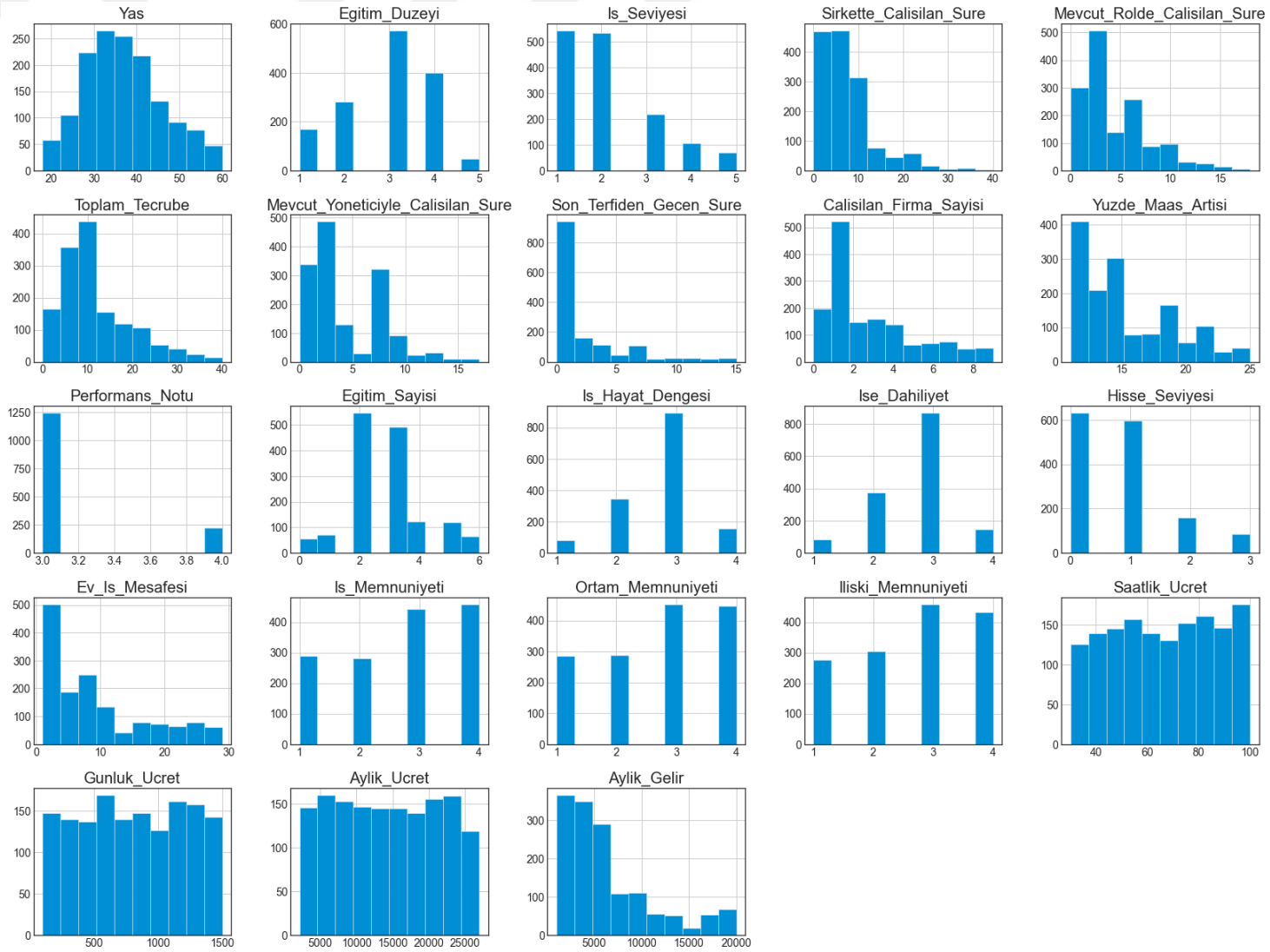
Tablo 13. Veri İstatistikleri

İstatistik	Calisan_No	Yas	Egitim_Duzeyi	Calisan_Sayisi	Is_Seviyesi	Sirkette_Calisilan_Sure	Mevcut_Rolde_Calisilan_Sure
count	1470.00	1470.00	1470.00	1470.00	1470.00	1470.00	1470.00
mean	1024.87	36.92	2.91	1.00	2.06	7.01	4.23
std	602.02	9.14	1.02	0.00	1.11	6.13	3.62
min	1.00	18.00	1.00	1.00	1.00	0.00	0.00
25%	491.25	30.00	2.00	1.00	1.00	3.00	2.00
50%	1020.50	36.00	3.00	1.00	2.00	5.00	3.00
75%	1555.75	43.00	4.00	1.00	3.00	9.00	7.00
max	2068.00	60.00	5.00	1.00	5.00	40.00	18.00

İstatistik	Toplam_Tecrube	Mevcut_Yoneticiyle_Calisilan_Sure	Son_Terfiden_Gecen_Sure	Calisilan_Firma_Sayisi	Yuzde_Maas_Artisi	Performans_Notu	Egitim_Sayisi
count	1470.00	1470.00	1470.00	1470.00	1470.00	1470.00	1470.00
mean	11.28	4.12	2.19	2.69	15.21	3.15	2.80
std	7.78	3.57	3.22	2.50	3.66	0.36	1.29
min	0.00	0.00	0.00	0.00	11.00	3.00	0.00
25%	6.00	2.00	0.00	1.00	12.00	3.00	2.00
50%	10.00	3.00	1.00	2.00	14.00	3.00	3.00
75%	15.00	7.00	3.00	4.00	18.00	3.00	3.00
max	40.00	17.00	15.00	9.00	25.00	4.00	6.00

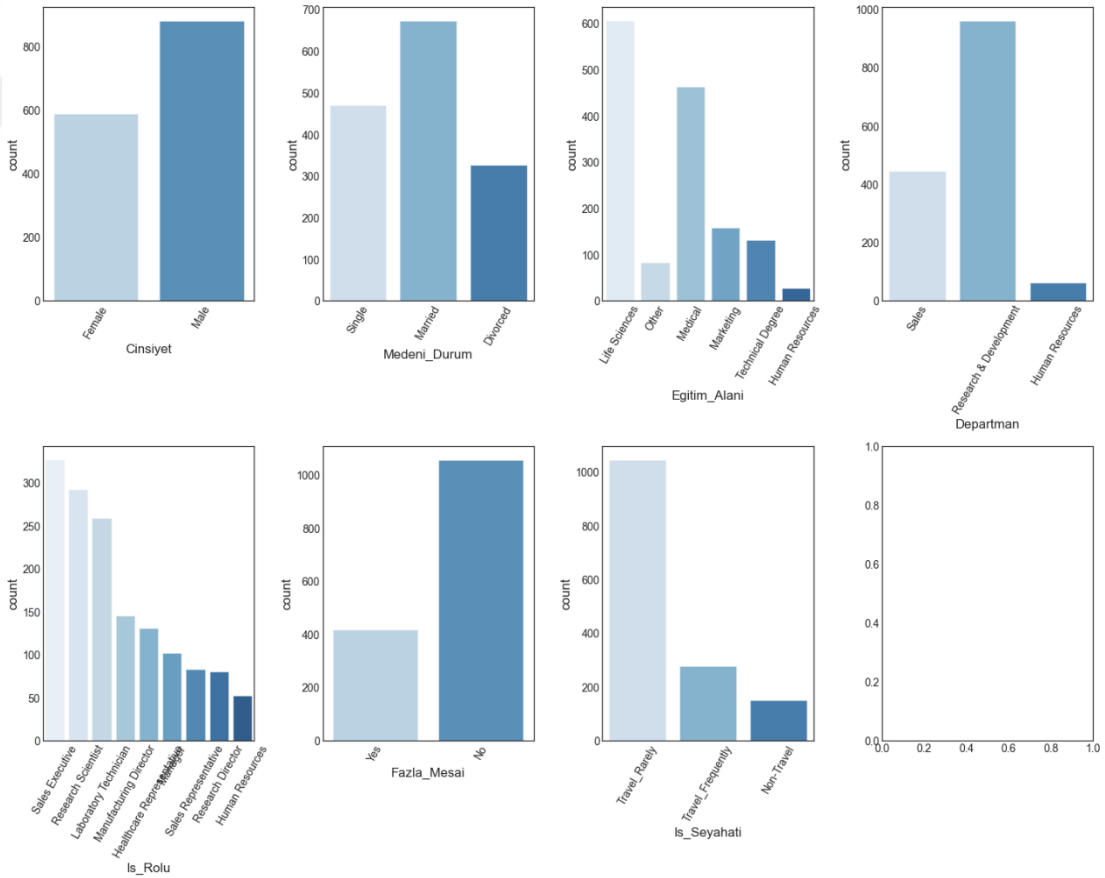
İstatistik	Is_Hayat_Dengesi	Ise_Dahiliyet	Standart_Saatler	Hisse_Seviyesi	Ev_Is_Mesafesi	Is_Memnuniyeti	Ortam_Memnuniyeti
count	1470.00	1470.00	1470.00	1470.00	1470.00	1470.00	1470.00
mean	2.76	2.73	80.00	0.79	9.19	2.73	2.72
std	0.71	0.71	0.00	0.85	8.11	1.10	1.09
min	1.00	1.00	80.00	0.00	1.00	1.00	1.00
25%	2.00	2.00	80.00	0.00	2.00	2.00	2.00
50%	3.00	3.00	80.00	1.00	7.00	3.00	3.00
75%	3.00	3.00	80.00	1.00	14.00	4.00	4.00
max	4.00	4.00	80.00	3.00	29.00	4.00	4.00

İstatistik	Iliski_Memnuniyeti	Saatlik_Ucret	Gunluk_Ucret	Aylik_Ucret	Aylik_Gelir
count	1470.00	1470.00	1470.00	1470.00	1470.00
mean	2.71	65.89	802.49	14313.10	6502.93
std	1.08	20.33	403.51	7117.79	4707.96
min	1.00	30.00	102.00	2094.00	1009.00
25%	2.00	48.00	465.00	8047.00	2911.00
50%	3.00	66.00	802.00	14235.50	4919.00
75%	4.00	83.75	1157.00	20461.50	8379.00
max	4.00	100.00	1499.00	26999.00	19999.00



Şekil 26. Numerik Veri Grafikleri

Sayısal değişkenler için yapılan bu analizler dışında kategorik değişkenlerin de incelenmesi gerekmektedir. Bu nedenle, her kategorik değişkenin aldığı değerlerin frekansları incelenmiş ve grafiği aşağıdaki şekilde çizilmiştir.

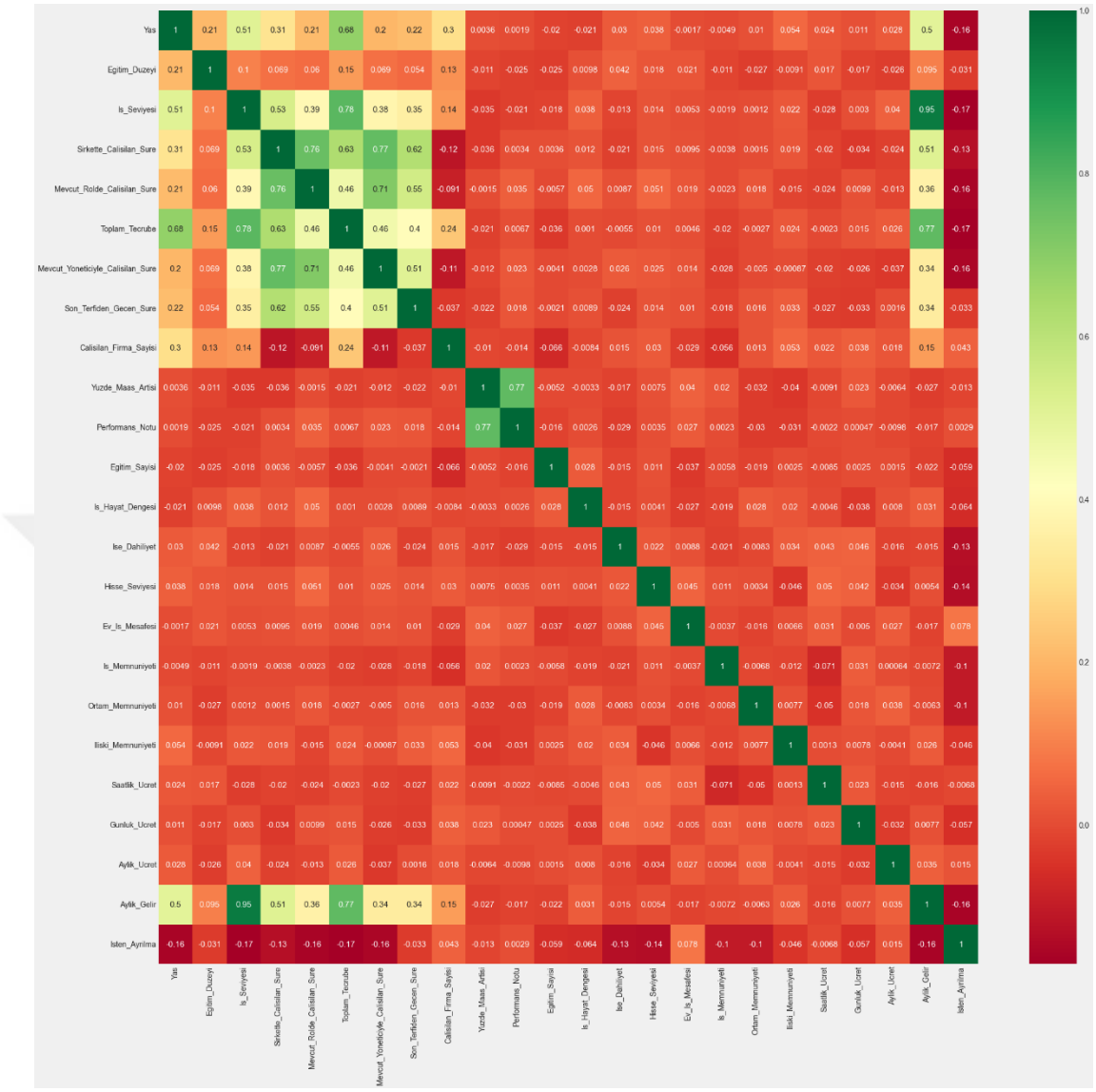


Şekil 27. Kategorik Veri Grafikleri

Kategorik değişkenlerin incelenmesi sonucu elde edilen bulgular şunlardır:

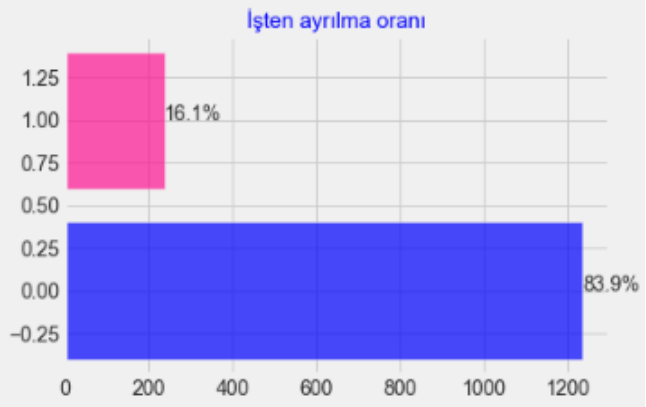
- Çalışanların %60'ı erkektir.
- Çalışanların %45'i evli, %32'si bekar ve gerisi boşanmıştır.
- Çalışanların %41'i sosyal bilimler, %32'si tıp/medikal, %11'i pazarlama, %9'u teknik bilimler ve %2'si insan kaynakları alanında eğitimlerini tamamlamıştır. Bunun dışında kalan kesim ise "Diğer" olarak sınıflandırılmıştır.
- Çalışanların %65'i Araştırma-Geliştirme (Arge) departmanında çalışmaktadır. Bunun dışında %30'u satış ve geri kalan %5'i insan kaynakları departmanlarındadır.
- Çalışanların işyerindeki rolleri hacim olarak sıralandığında satış yöneticisi, araştırma bilim insanı, laboratuvar teknisyeni, üretim direktörü, sağlık sorumlusu, müdür, satış sorumlusu, araştırma direktörü ve insan kaynakları sorumlusu şeklindedir.
- Çalışanların %72'si fazla mesai yapmamaktadır.
- Çalışanların %10'u hiç iş seyahati yapmamaktadır; %71'i nadiren seyahat etmektedir ve geri kalanlar ise sık sık seyahat etmektedir.

Değişkenlerin tekil incelenmesi dışında yapılabilecek bir başka analiz de değişkenler arasındaki korelasyonun incelenmesi olacaktır. Böylece, değişkenler arasındaki lineer ilişkinin düzeyi tespit edilebilir. Aşağıdaki tabloda kırmızı renkler korelasyon bulunmadığını, yeşil renkler ise korelasyonun olduğunu göstermektedir. Buna göre; iş seviyesi, şirkette çalışılan süre, mevcut rolde çalışılan süre, mevcut yönetici ile çalışılan süre, son terfiden geçen süre ve toplam tecrübe diğer değişkenlere göre kendi aralarında daha yüksek korelasyona sahiptir. Bunun dışında, iş seviyesi ile aylık gelir arasında da bekleneceği gibi yüksek bir korelasyon seviyesi mevcuttur.



Şekil 28. Korelasyon Matrisi

Değişkenlerin kendi aralarındaki korelasyona bakıldıktan sonra hedef değişken analizleri yapılmalıdır. 1470 çalışan üzerinden yapılan analizde işten ayrılma oranı %16.1 olarak görülmektedir.

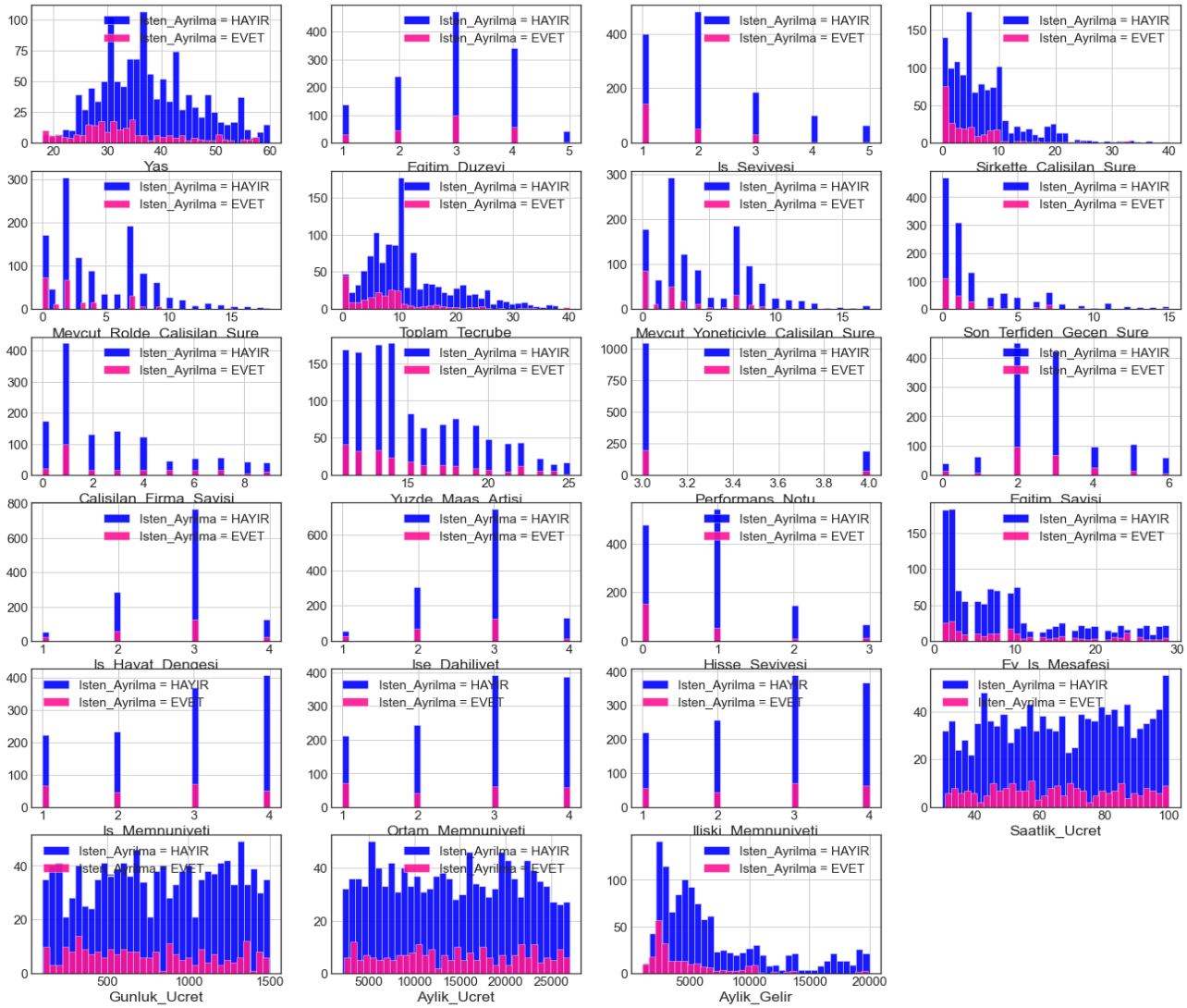


Şekil 29. Hedef Değişken Analizi

Hedef değişkenin dağılımına bakılmasının ardından hedef değişkenin, yani işten ayrılma durumunun, "Evet" veya "Hayır" değerleri bazında da değişkenlerin dağılımının tekrar incelenmesi de veri ile ilgili ek içgörüler elde edilmesini sağlayabilir. Aşağıda gösterilen bu sonuçlar incelendiğinde şu bulgular elde edilmektedir:

- İş seviyesi değişkeni incelendiğinde, 1 numaralı seviyedeki çalışanların daha yüksek oranda işten ayrıldığı gözlemlenebilir.
- Şirkette çalışılan süre kısaldıkça da işten ayrılma oranında bir artış olduğu görülmektedir. Bu da şirket bağlılığının henüz oluşmadığı durumda ayrılmaların yüksek olduğuna işaret ediyor olabilir.
- Mevcut rolde veya mevcut yöneticiyle çalışılan sürenin kısa olması da işten ayrılmalar ile bağlantılı görünüyor. Bunun açıklaması da başlanan role veya yeni çalışılmaya başlanan yöneticiye uyum sıkıntısı şeklinde yapılabilir.
- Toplam tecrübenin az olması da işten ayrılmalar ile ilişki olduğunu gösteriyor. Üst pozisyonlarda daha az ilan çıkması ve personelin tecrübesi arttıkça daha az değişiklik istemesi de bunu açıklamaktadır.
- Çalışılan firma sayısı arttıkça işten ayrılma durumu da artış göstermektedir. Bu da iş değiştirme alışkanlığı olduğunu göstermektedir.
- Yüzde maaş artışı veya performans notunun bu grafiklere göre işten ayrılmaya etkisi anlaşılamamaktadır.
- İş, ortam veya ilişki memnuniyeti azaldıkça da işten ayrılma oranının arttığı görülmektedir.
- Saatlik, günlük veya aylık ücret ile işten ayrılma arasında bir bağlantı da bu grafikler üzerinden anlaşılamamaktadır.

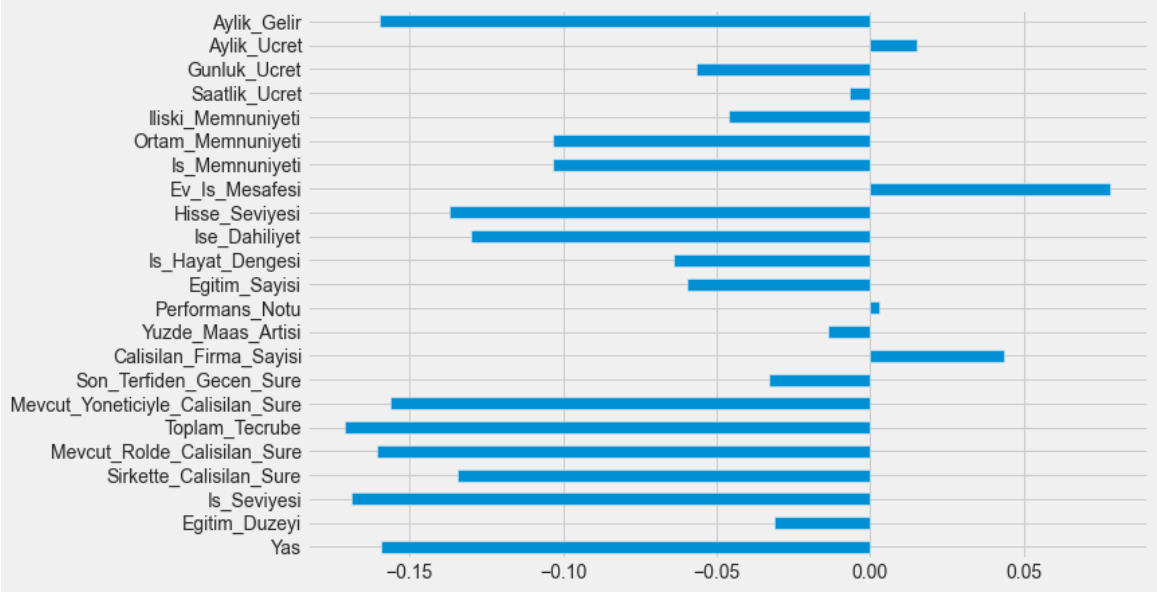
- Medeni durum ile işten ayrılmalar arasında açık bir ilişki vardır ve bekar çalışanların işten ayrılma oranı daha yüksektir.
- Satış departmanındaki çalışanlar diğer departmanlara göre daha yüksek oranda iş değiştirmektedir. Bu da piyasada bu alanda daha fazla açık pozisyon olması, bu alanda daha kolay iş değiştirilebilmesi veya departmandaki olası bir memnuniyetsizlik ile açıklanabilir.
- Roller incelendiğinde ise laboratuvar teknisyenlerinin işten ayrılma oranının diğer rollere göre daha yüksek olduğu görülmektedir.
- Fazla mesai yapmanın işten ayrılma üzerinde bariz bir etkisi vardır ve fazla mesai yapan çalışanlarda işten ayrılma oranı yüksektir.
- İş seyahati yapan çalışanların işten ayrılma durumlarında bir etki gözlemlenmektedir.



Şekil 30. İşten Ayrılma Durumuna Göre Veri Dağılımları

İşten ayrılma durumu ile her bir değişkenin ilişkisini incelemenin bir diğer yolu da bu bilgiler arasındaki korelasyonun incelenmesidir. Aşağıda, verideki tüm sayısal değişkenlerin işten ayrılma durumu ile arasındaki doğrusal ilişki düzeyini görmek amacıyla korelasyon incelemesi yapılmıştır. Buna göre bazı temel bulgular aşağıda verilmiştir ve bir önceki bulgular ile benzer olduğu söylenebilir.

- Aylık gelir ile işten ayrılma arasında ters ilişki söz konusudur
- Memnuniyet ile işten ayrılma arasında ters ilişki söz konusudur
- Hisse seviyesi ile işten ayrılma arasında ters ilişki söz konusudur
- İşe dahiliyet ile işten ayrılma arasında ters ilişki söz konusudur
- İş hayat dengesi ile işten ayrılma arasında ters ilişki söz konusudur
- Eğitim sayısı ile işten ayrılma arasında ters ilişki söz konusudur
- Yüzde maaş artışı ile işten ayrılma arasında ters ilişki söz konusudur
- Mevcut yöneticiyle çalışıla süre, mevcut rolde çalışılan süre, toplam tecrübe gibi değişkenler ile işten ayrılma arasında ters ilişki söz konusudur
- Yaş ile işten ayrılma arasında ters ilişki söz konusudur
- İş seviyesi ile işten ayrılma arasında ters ilişki söz konusudur
- Ev-iş mesafesi ile işten ayrılma arasında doğru ilişki söz konusudur
- Çalışılan firma sayısı ile işten ayrılma arasında ters ilişki söz konusudur
- Performans notunun işten ayrılma üzerindeki doğrusal etkisi yok sayılabilir



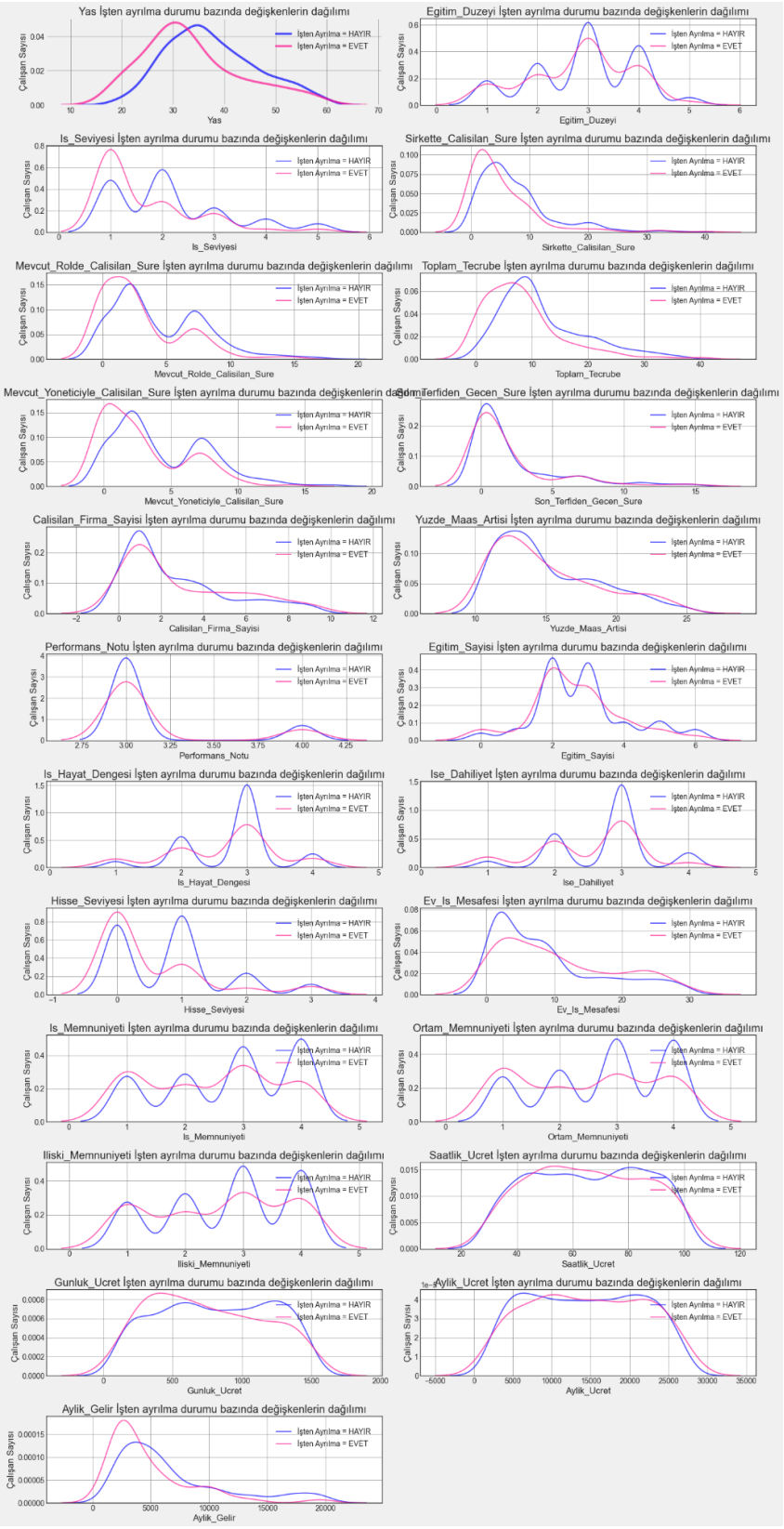
Şekil 31. Sayısal Değişkenler ile Hedef Değişken Arası Doğrusal İlişki

İşten ayrılma durumu bazında her bir değişkenin ortalama ve standart sapmalarını incelemek de bir başka analiz yöntemidir. Bu analiz sonucunda diğer analizlerde olduğu gibi hangi değişkenlerin tekil olarak etkili olup olmadığı ile ilgili yorum yapılabilmektedir. Bu tablo aşağıda görülebilir.

Tablo 14. İşten Ayrılma Durumuna Göre İstatistikler

	İşten Ayrılma = HAYIR		İşten Ayrılma = EVET		
	Ortalama	**Std. Sapma**	**Ortalama**	**Std. Sapma**	**Etki Yönü**
Yaş	37.6	8.9	33.6	9.7	-
Eğitim Düzeyi	2.9	1.0	2.8	1.0	-
İş Seviyesi	2.1	1.1	1.6	0.9	-
Şirkette Çalışılan Süre	7.4	6.1	5.1	5.9	-
Mevcut Rolde Çalışılan Süre	4.5	3.6	2.9	3.2	-
Toplam Tecrübe	11.9	7.8	8.2	7.2	-
Mevcut Yöneticiyle Çalışılan Sü	4.4	2.6	2.9	3.1	-
Son Terfiden Geçen Süre	2.2	3.2	1.9	3.1	-
Çalışılan Firma Sayısı	2.6	2.5	2.9	2.7	+
Yüzde Maaş Artışı	15.2	3.6	15.1	3.8	-
Performans Notu	3.2	0.4	3.2	0.4	=
Eğitim Sayısı	2.8	1.3	2.6	1.3	-
İş Hayat Dengesi	2.8	0.7	2.7	0.8	-
İşe Dahiliyet	2.8	0.7	2.5	0.8	-
Hisse Seviyesi	0.8	0.8	0.5	0.9	-
Ev-İş Mesafesi	8.9	8.0	10.6	8.4	+
İş Memnuniyeti	2.8	1.1	2.5	1.1	-
Ortam Memnuniyeti	2.8	1.1	2.5	1.2	-
İlişki Memnuniyeti	2.7	1.1	2.6	1.1	-
Saatlik Ücret	66.0	20.4	65.6	20.1	-
Günlük Ücret	812.5	403.0	750.4	401.1	-
Aylık Ücret	14,265.8	7,099.4	14,559.3	7,192.9	+
Aylık Gelir	6,832.7	4,816.3	4,787.1	3,632.5	-

Yine benzer bir analiz için değişkenlerin çizgi grafiklerinin çizilmesi de benzer bulgulara ulaşmamıza olanak sağlamaktadır.



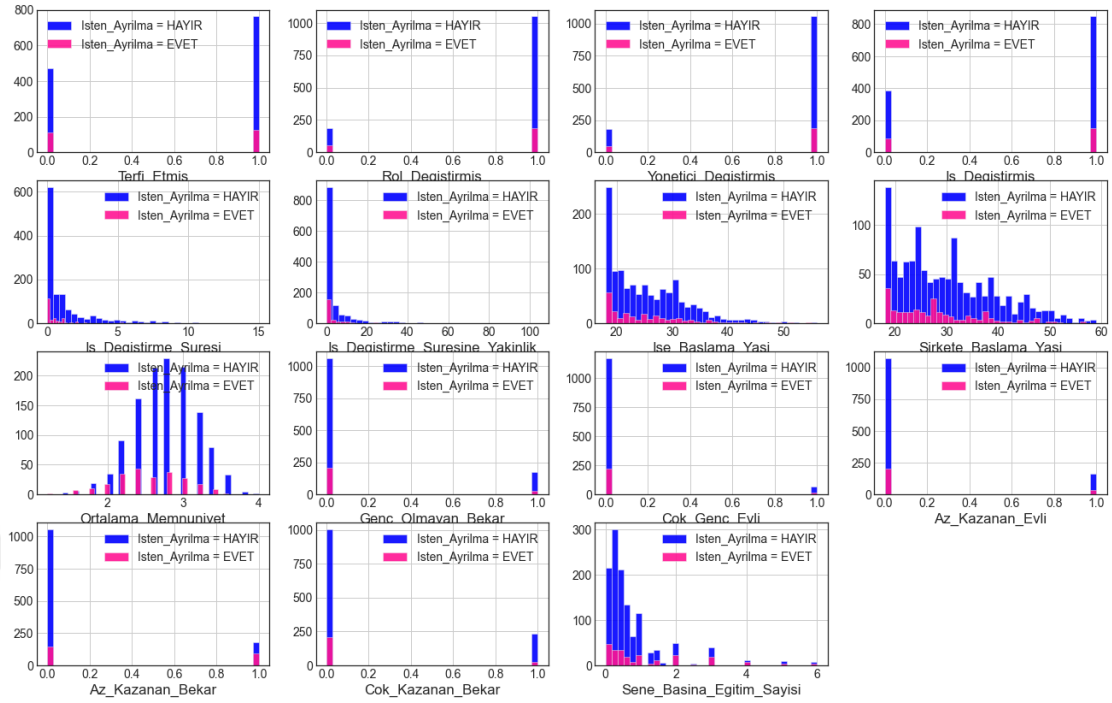
Şekil 32. İşten Ayrılma Durumu Bazında Dağılımlar

Tüm bu analiz yöntemleriyle birlikte veri analizi yapılmış diyebiliriz. Ancak bundan sonra modellemede kullanılabilecek ek verilerin türetilmesi de anlamlı olacaktır. Veriyi zenginleştirme amacıyla aşağıdaki değişkenler türetilmiştir:

Tablo 15. Yeni Türetilen Değişkenler

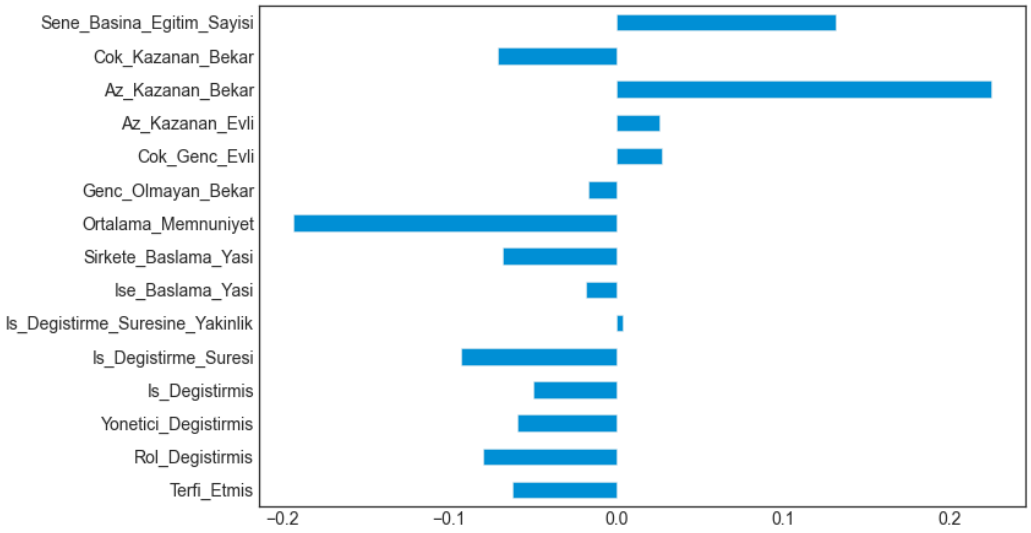
Yeni Değişken	Formül
Terfi Etmiş	data['Son_Terfiden_Gecen_Sure'] > 0
Rol Değiştirmiş	data['Mevcut_Rolde_Calisilan_Sure'] < data['Sirkette_Calisilan_Sure']
Yönetici Değiştirmiş	data['Mevcut_Yoneticiyle_Calisilan_Sure'] < data['Sirkette_Calisilan_Sure']
İş Değiştirmiş	data['Sirkette_Calisilan_Sure'] < data['Toplam_Tecrube']
İş Değiştirme Süresi	(data['Toplam_Tecrube'] - data['Sirkette_Calisilan_Sure'])/(data['Calisilan_Firma_Sayisi'])
İş Değiştirme Süresine Yakınlık	data['Sirkette_Calisilan_Sure']/data['Is_Degistirme_Suresi']
İlk İşe Başlama Yaşı	data['Yas'] - data['Toplam_Tecrube']
Şirkete Başlama Yaşı	data['Yas'] - data['Sirkette_Calisilan_Sure']
Ortalama Memnuniyet	data['Ise_Dahiliyet'] + data['Is_Memnuniyeti'] + data['Ortam_Memnuniyeti'] + data['Iliski_Memnuniyeti'] + data['Is_Hayat_Dengesi'])/5
Genç Olmayan Bekar	data['Yas'] > 35 and data['Medeni_Durum'] == 'Single'
Çok Genç Evli	data['Yas'] < 28 and data['Medeni_Durum'] == 'Married'
Az Kazanan Evli	data['Aylik_Gelir'] < 3500 and data['Medeni_Durum'] == 'Married'
Az Kazanan Bekar	data['Aylik_Gelir'] <3500 and (data['Medeni_Durum'] == 'Single' or data['Medeni_Durum'] == 'Divorced')
Çok Kazanan Bekar	data['Aylik_Gelir'] > 6500 and (data['Medeni_Durum'] == 'Single' or data['Medeni_Durum'] == 'Divorced')
Sene Başına Eğitim Sayısı	data['Egitim_Sayisi']/data['Sirkette_Calisilan_Sure']

Yukarıda ismi geçen ve yeni türetilmiş olan değişkenler de veri içerisinde hazır olarak yer verilen değişkenler gibi analiz edilmiştir. Aşağıdaki grafikler yeni türetilen değişkenlerin hedef değişken ile ilişkisini ortaya çıkarmaktadır.



Şekil 33. Türetilen Değişkenlerin İşten Ayrılma Verisine Göre Dağılımı

Yeni türetilen değişkenlerin hedef değişken ile arasındaki korelasyon daha önceki değişkenlere göre daha düşük bir aralıkta dağılmaktadır.



Şekil 34. Türetilen Değişkenler ile Hedef Değişken Arasındaki İlişki

İşten ayrılma durumu bazında her yeni değişkenin ortalama ve standart sapmaları incelendiğinde hangi değişkenlerin tekil olarak etkili olup olmadığı ile ilgili yorum yapılabilmektedir.

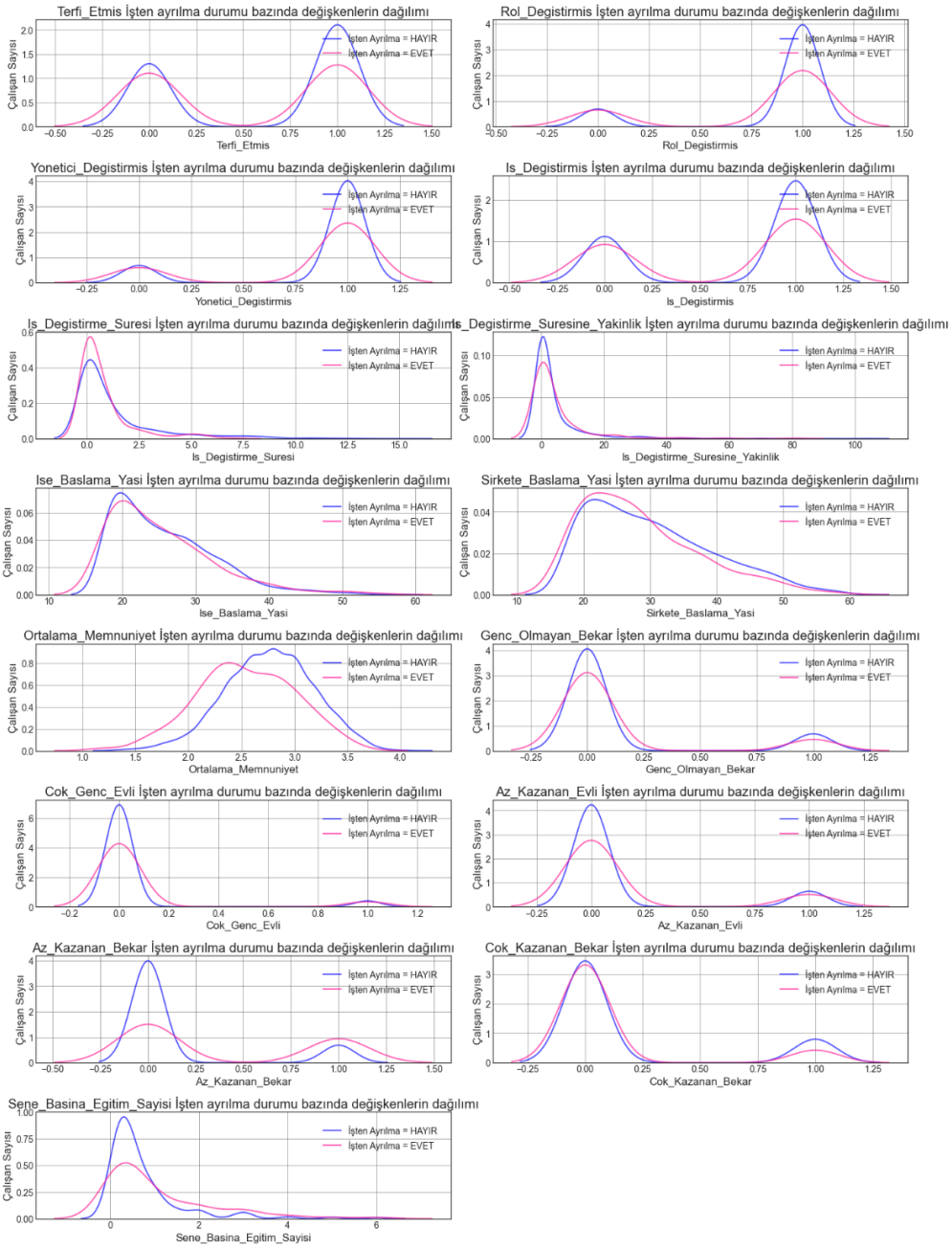
Buna göre, değişkenler arasındaki farklar minimal görünse de evli-bekar olma durumları veya sene başına düşen eğitim sayısı fark yaratan değişkenlere benzemektedir.

Tablo 16. Türetilen Değişkenlerin İşten Ayrılmaya Göre İstatistikleri

	İşten Ayrılma = HAYIR		İşten Ayrılma = EVET		
	Ortalama	**Std. Sapma**	**Ortalama**	**Std. Sapma**	**Etki Yönü**
Terfi Etmiş	0.6	0.5	0.5	0.5	-
Rol Değiştirmiş	0.9	0.4	0.8	0.4	-
Yönetici Değiştirmiş	0.9	0.4	0.8	0.4	-
İş Değiştirmiş	0.7	0.5	0.6	0.5	-
İş Değiştirme Süresi	1.3	2.2	0.8	2.2	-
İş Değiştirme Süresi Yakınlık	4.2	9.6	4.3	9.4	+
İşe Başlama Yaşı	25.7	6.8	25.4	7.2	-
Şirkete Başlama Yaşı	30.2	9.3	28.5	8.8	-
Ortalama Memnuniyet	2.8	0.4	2.5	0.5	-
Genç Olmayan Bekar	0.1	0.3	0.1	0.3	=
Çok Genç Evli	0.1	0.2	0.1	0.3	=
Az Kazanan Evli	0.1	0.3	0.2	0.4	+
Az Kazanan Bekar	0.1	0.4	0.4	0.5	+
Çok Kazanan Bekar	0.2	0.4	0.1	0.3	-
Sene Başına Eğitim Sayısı	0.7	0.9	1.1	1.2	+

Çizgi grafikler ile de yeni türetilen değişkenlerin hedef değişken ile ilişkisi farklı bir boyuttan incelenmiştir.

Tablo 17. İşten Ayrılma Durumu Bazında Türetilmiş Veri Dağılımları



Şu ana kadar yapılmış tüm analizler tekli faktör analizleri olup bir sonraki adımda model algoritmaları yardımıyla çoklu faktör analizleri yapılmış olacaktır.

5.4. Modelleme

Veride yapılan ek alan türetme işlemlerinden sonra birden fazla modelleme algoritması kullanılarak modelleme yapılmıştır. Modelleme için kullanılacak her algoritma ayrı bir alt başlık altında detaylandırılacak olup şu algoritmalar kullanılmıştır:

- Karar ağacı
- Lojistik regresyon
- Ekstrem gradyan artırma
- Rassal ağaç
- Catboost
- Light GBM

Modelleme öncesinde verinin %70'i model geliştirme ve %30'u test için rassal şekilde ayrılmıştır.

5.4.1. Karar Ağacı Model

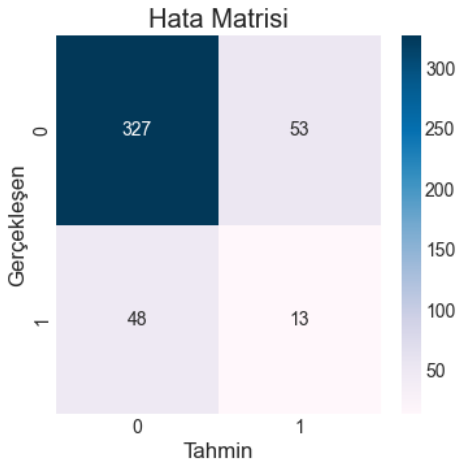
Karar ağacı algoritması kullanılarak geliştirilen modelin performans metrikleri şu şekildedir:

Tablo 18. Karar ağacı modeli performans metrikleri

Kitle	Doğruluk	Kesinlik	Duyarlılık	F1 Skoru
Geliştirme	1	1	1	1
Test	0.7710	0.7786	0.7710	0.7747

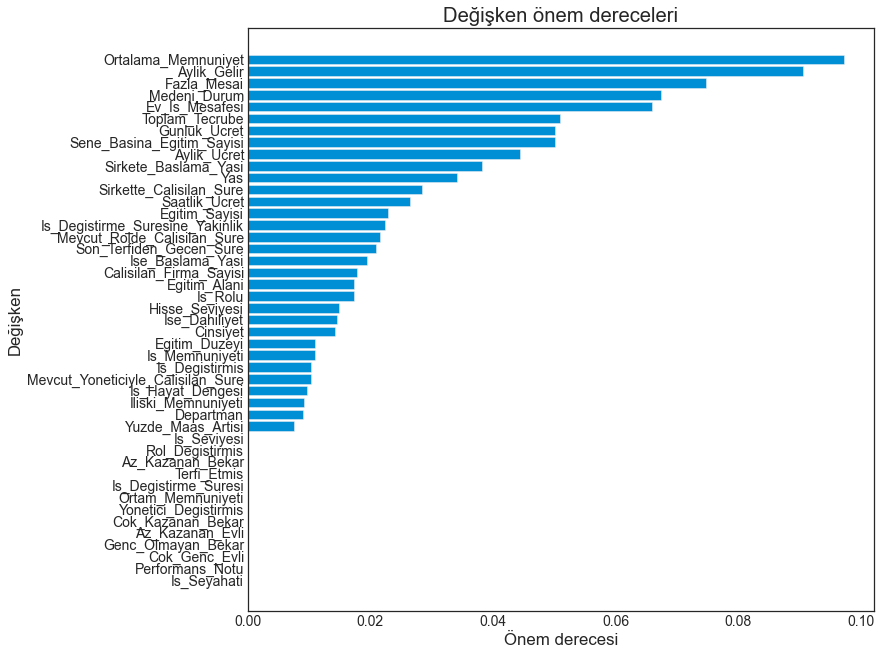
Yukarıdaki performans tablosuna göre model yüksek oranda aşırı öğrenme (overfit) göstermiştir. Hata matrisine bakıldığında,

- İşten ayrılmayanların (327+53 kişi) %86'sı model tarafından da işten ayrılmayacak şekilde tahmin edilmiştir. İşten ayrılmayan ve modelin de ayrılmayacakları yönünde tahminde bulunduğu %86'lık kesim "doğru pozitif" olarak isimlendirilmektedir.
- 380 kişinin %14'ünün işten ayrılmadığı halde işten ayrılacağı düşünülmüştür. Bu %14'lük kitle "yanlış negatif" olarak isimlendirilmektedir.
- Benzer şekilde işten ayrılan 61 kişinin 13'ünün işten ayrılacağı bilinebilmiştir ve bu kitle "doğru negatif" olarak isimlendirilmektedir.
- 61 kişiden 48'i ise işten ayrılmış olduğu halde işten ayrılmayacağı tahmini yapılmıştır. Bu kitle de "yanlış pozitif" olarak isimlendirilmektedir.



Şekil 35. Karar ağacı modeli hata matrisi

Modelde anlamlı çıkan değişkenler önem dereceleri ile aşağıdaki şekilde sıralanmıştır.



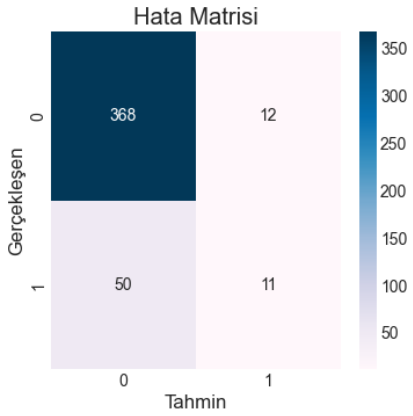
Şekil 36. Karar ağacı modeli değişken önem dereceleri

5.4.2. Karar Ağacı Modeli – Hiperparametre Optimizasyonu ile

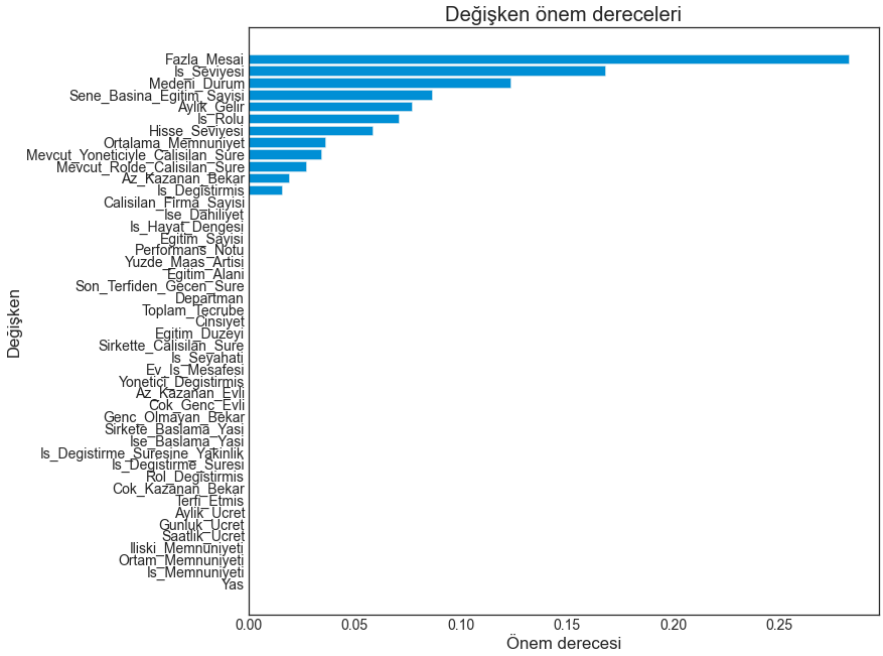
Karar ağacı algoritması kullanılıp bir de hiperparametre optimizasyonu yapıldığında geliştirilen modelin performans metrikleri ve önemli değişkenleri aşağıdaki şekildedir:

Tablo 19. Karar ağacı modeli performans metrikleri (2)

Kitle	Doğruluk	Kesinlik	Duyarlılık	F1 Skoru
Geliştirme	0.8630	0.8476	0.8630	0.8473
Test	0.8594	0.8248	0.8594	0.8310



Şekil 37. Karar ağacı modeli hata matrisi (2)



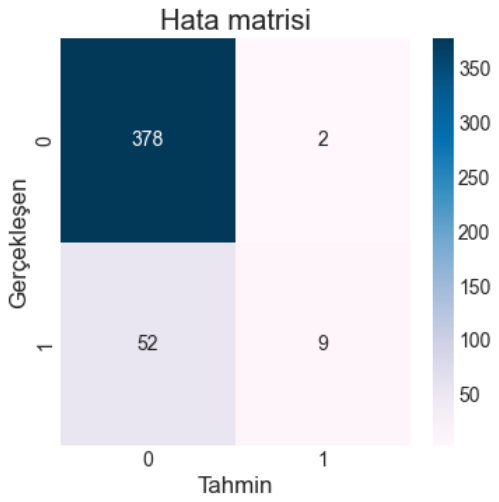
Şekil 38. Karar ağacı modeli değişken önem dereceleri (2)

5.4.3. Rassal Ağaç Modeli

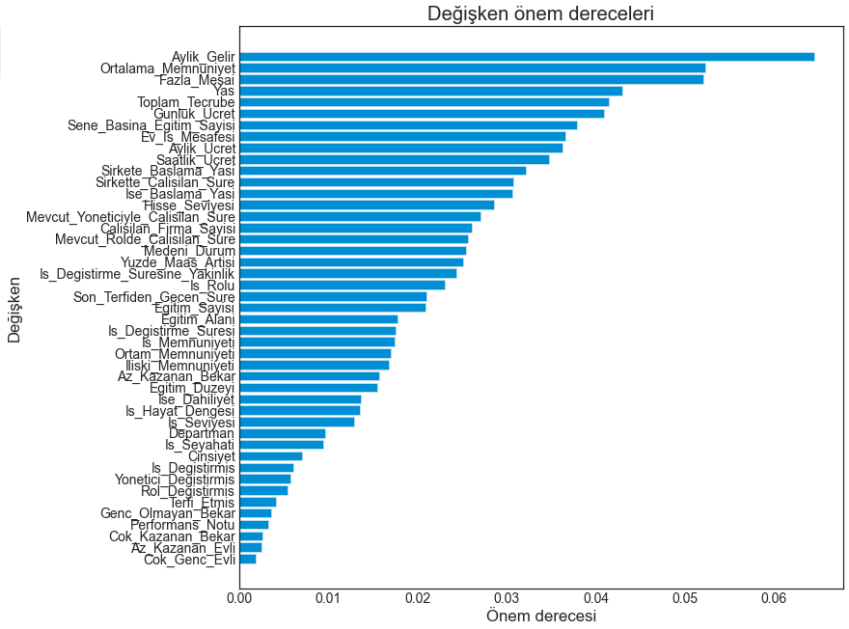
Rassal ağaç algoritması kullanılarak geliştirilen modelin performans metrikleri ve önemli değişkenleri aşağıdaki şekildedir:

Tablo 20. Rassal ağaç modeli performans metrikleri

Kitle	Doğruluk	Kesinlik	Duyarlılık	F1 Skoru
Geliştirme	1	1	1	1
Test	0.8776	0.8706	0.8776	0.8388



Şekil 39. Rassal ağaç modeli hata matrisi



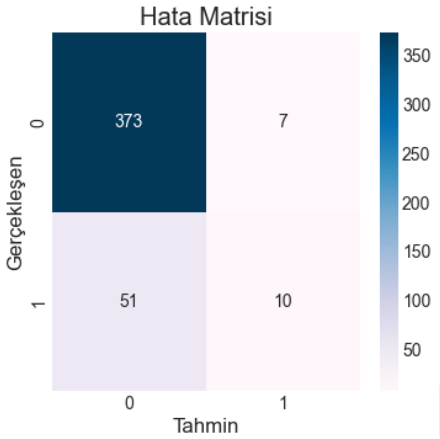
Şekil 40. Rassal ağaç modeli değişken önem dereceleri

5.4.4. Rassal Ağaç Modeli - Hiperparametre Optimizasyonu ile

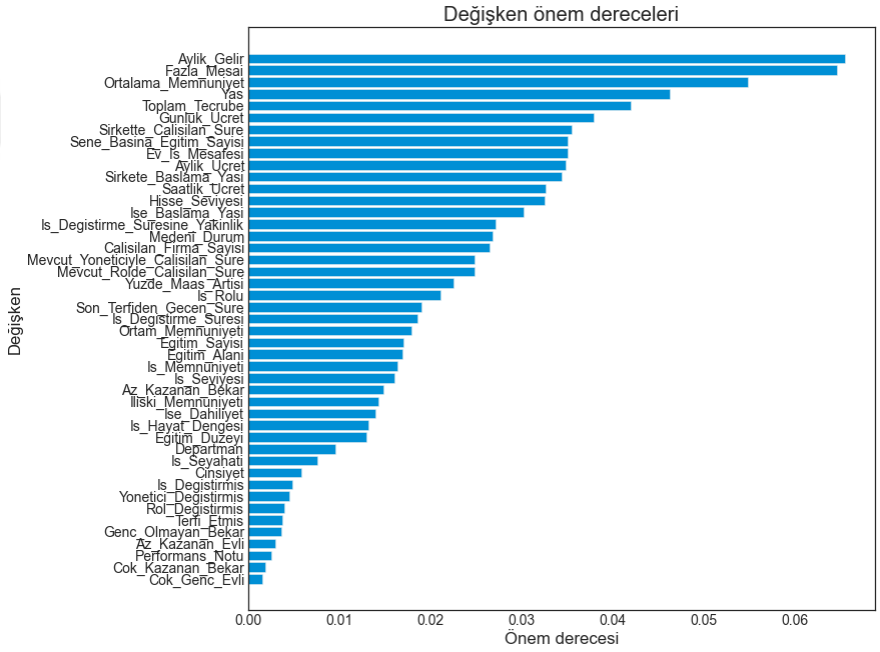
Rassal ağaç algoritması kullanılarak geliştirilen modelin performans metrikleri ve önemli değişkenleri aşağıdaki şekildedir:

Tablo 21. Rassal Ağaç Modeli Performans Metrikleri (2)

Kitle	Doğruluk	Kesinlik	Duyarlılık	F1 Skoru
Geliştirme	1	1	1	1
Test	0.8685	0.8394	0.8685	0.8350



Şekil 41. Rassal ağaç modeli hata matrisi (2)



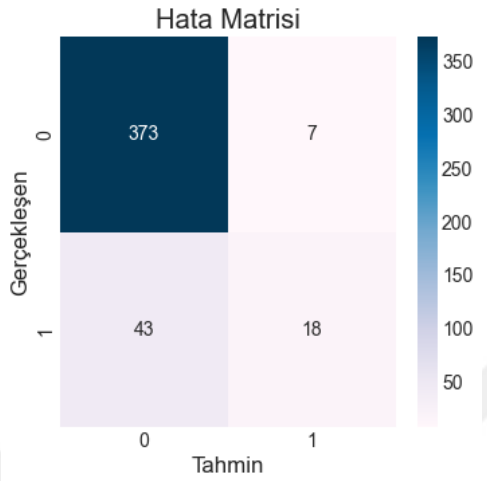
Şekil 42. Rassal ağaç modeli değişken önem dereceleri (2)

5.4.5. Lojistik Regresyon Modeli

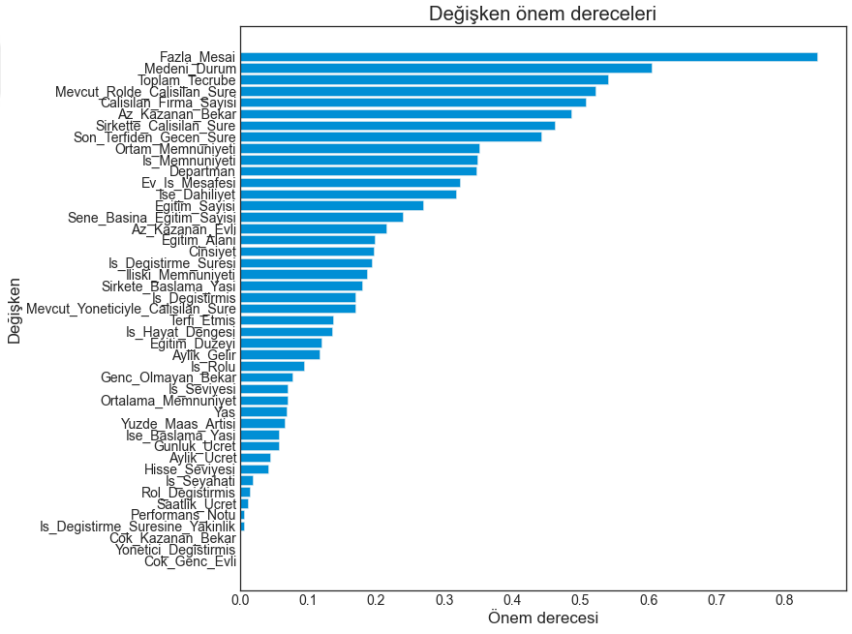
Lojistik regresyon algoritması kullanılarak geliştirilen modelin performans metrikleri ve önemli değişkenleri aşağıdaki şekildedir:

Tablo 22. Lojistik Regresyon Modeli Performans Metrikleri

Kitle	Doğruluk	Kesinlik	Duyarlılık	F1 Skoru
Geliştirme	0.8853	0.8769	0.8853	0.8734
Test	0.8866	0.8722	0.8866	0.8655



Şekil 43. Lojistik regresyon modeli hata matrisi



Şekil 44. Lojistik regresyon modeli değişken önem dereceleri

5.4.6. Destek Vektör Makinesi Modeli

Lojistik regresyon algoritması kullanılarak geliştirilen modelin performans metrikleri aşağıdaki şekildedir:

Tablo 23. Destek Vektör Modeli Performans Metrikleri

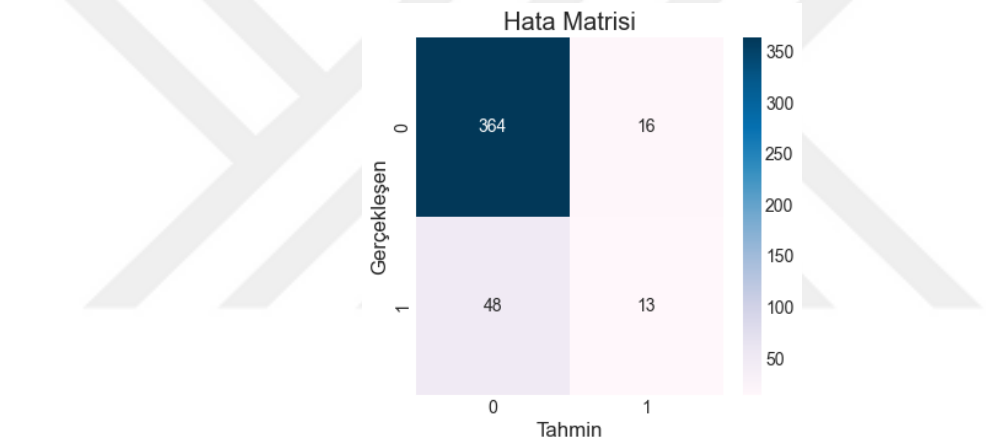
Kitle	Doğruluk	Kesinlik	Duyarlılık	F1 Skoru
Geliştirme	0.8552	0.8416	0.8552	0.8221
Test	0.8549	0.8058	0.8549	0.8151

5.4.7. Ekstrem Gradyan Artırma Modeli

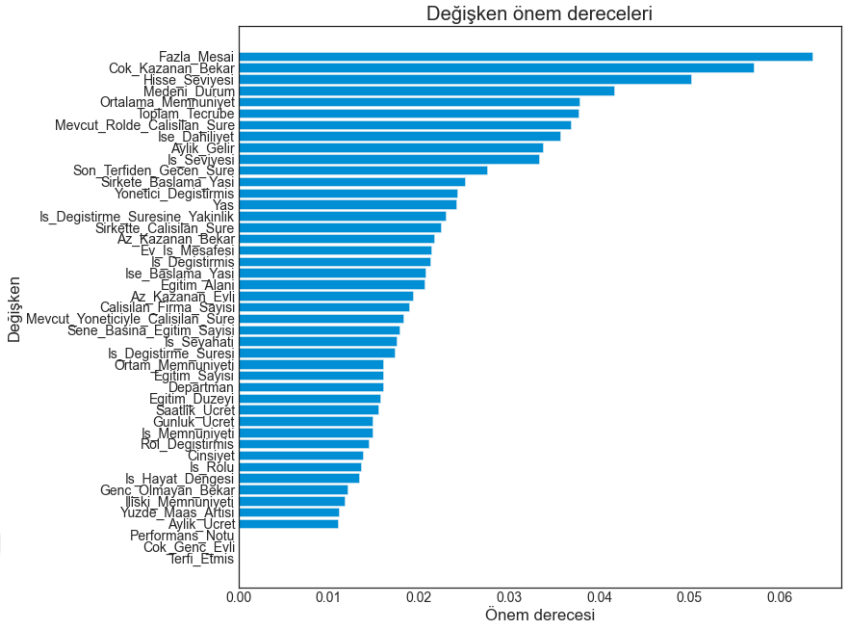
XGBoost algoritması kullanılarak geliştirilen modelin performans metrikleri ve önemli değişkenleri aşağıdaki şekildedir:

Tablo 24. Ekstrem Gradyan Artırma Modeli Performans Metrikleri

Kitle	Doğruluk	Kesinlik	Duyarlılık	F1 Skoru
Geliştirme	1	1	1	1
Test	0.8549	0.8233	0.8549	0.8320



Şekil 45. Ekstrem Gradyan Artırma modeli hata matrisi



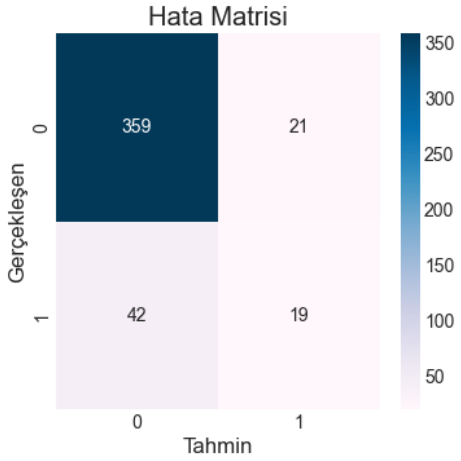
Şekil 46. Ekstrem gradyan artırma modeli değişken önem dereceleri

5.4.8. Uyarlanabilir Yükseltme Modeli

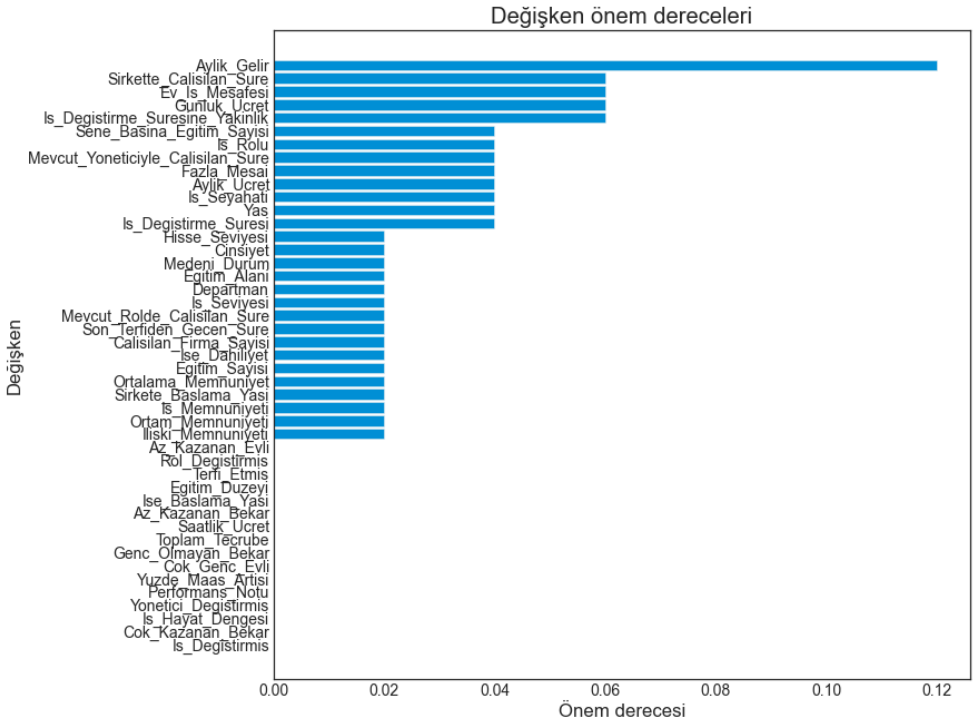
AdaBoost algoritması kullanılarak geliştirilen modelin performans metrikleri ve önemli değişkenleri aşağıdaki şekildedir:

Tablo 25. Uyarlanabilir Yükseltme Modeli Performans Metrikleri

Kitle	Doğruluk	Kesinlik	Duyarlılık	F1 Skoru
Geliştirme	0.9018	0.8963	0.9018	0.8943
Test	0.8571	0.8371	0.8571	0.8442



Şekil 47. Uyarlanabilir Yükseltme modeli hata matrisi



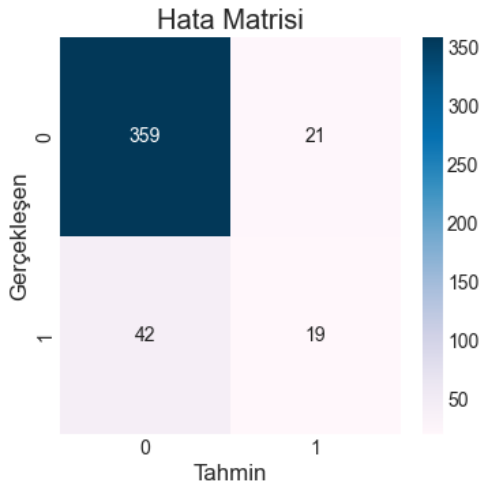
Şekil 48. Uyarlanabilir Yükseltme modeli değişken önem dereceleri

5.4.9. Catboost Modeli

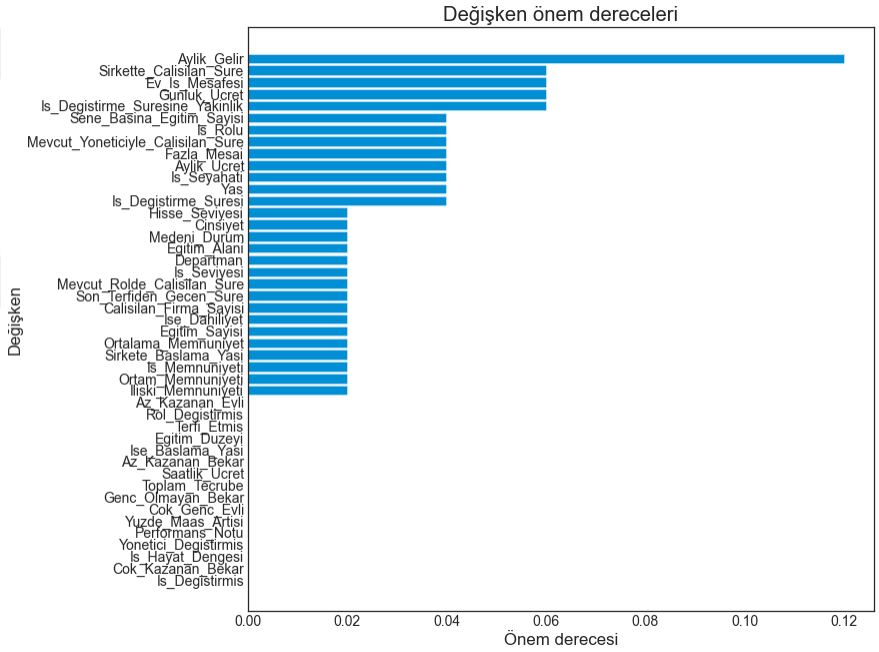
Catboost algoritması kullanılarak geliştirilen modelin performans metrikleri ve önemli değişkenleri aşağıdaki şekildedir:

Tablo 26. Catboost Modeli Performans Metrikleri

Kitle	Doğruluk	Kesinlik	Duyarlılık	F1 Skoru
Geliştirme	0.9883	0.9885	0.9883	0.9882
Test	0.8730	0.8507	0.8730	0.8382



Şekil 49. Catboost modeli hata matrisi



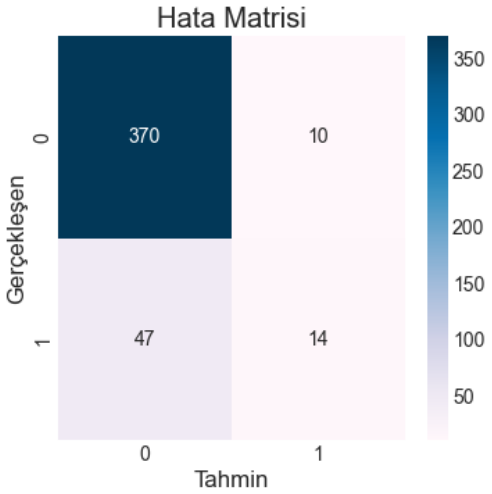
Şekil 50. Catboost modeli değişken önem dereceleri

5.4.10. Light GBM Modeli

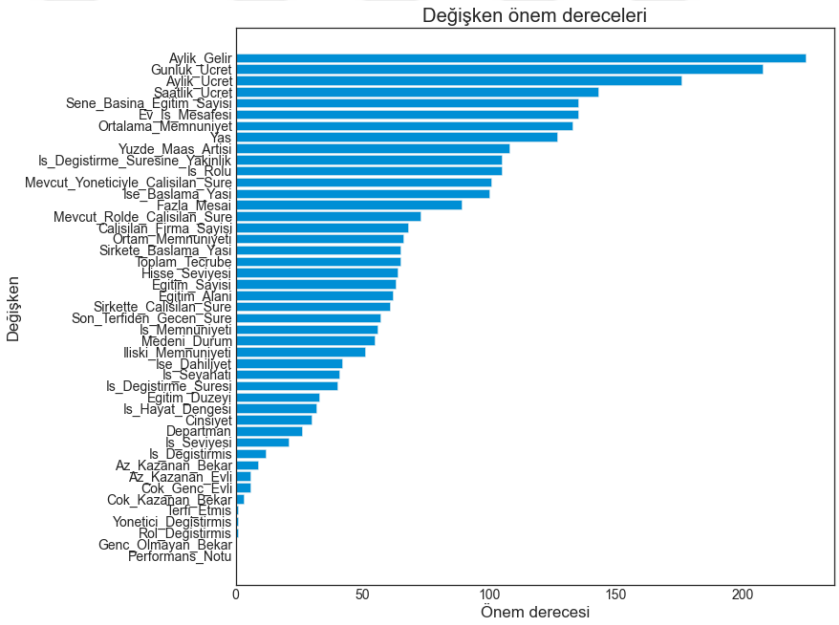
Light GBM algoritması kullanılarak geliştirilen modelin performans metrikleri ve önemli değişkenleri aşağıdaki şekildedir:

Tablo 27. Light GBM Modeli Performans Metrikleri

Kitle	Doğruluk	Kesinlik	Duyarlılık	F1 Skoru
Geliştirme	1	1	1	1
Test	0.8707	0.8452	0.8707	0.8456



Şekil 51. Light GBM modeli hata matrisi



Şekil 52. Light GBM modeli değişken önem dereceleri

5.4.11. Model Sonuçları Karşılaştırma

Farklı algoritmalar kullanılarak geliştirilmiş modellerin performanslarının karşılaştırılması için aşağıdaki şekilde bir tablo kullanılmıştır. Performanslar karşılaştırıldığında hem en iyi performansa sahip olan hem de aşırı öğrenme/uyum gerçekleşmemiş modelleme algoritması lojistik regresyon olduğundan bu model ile ilerlenmesine karar verilmiştir.

Tablo 28. Model Performans Karşılaştırması

Algoritma	Test Performansı			
	Doğruluk	Kesinlik	Duyarlılık	F1 Skoru
Karar Ağacı	0.7710	0.7786	0.7710	0.7747
Rassal Ağaç	0.8776	0.8706	0.8776	0.8388
Lojistik Regresyon	0.8866	0.8722	0.8866	0.8655
Destek Vektör Makinesi	0.8549	0.8058	0.8549	0.8151
Ekstrem Gradyan Artırma	0.8549	0.8233	0.8549	0.8320
Uyarlanabilir Yükseltme	0.8571	0.8371	0.8571	0.8442
Catboost	0.8730	0.8507	0.8730	0.8382
Light GBM	0.8707	0.8452	0.8707	0.8456

Açıklanabilirliği en kolay model olan lojistik regresyon modelinin değişkenleri şu şekildedir:

Tablo 29. Lojistik Regresyon Değişkenleri

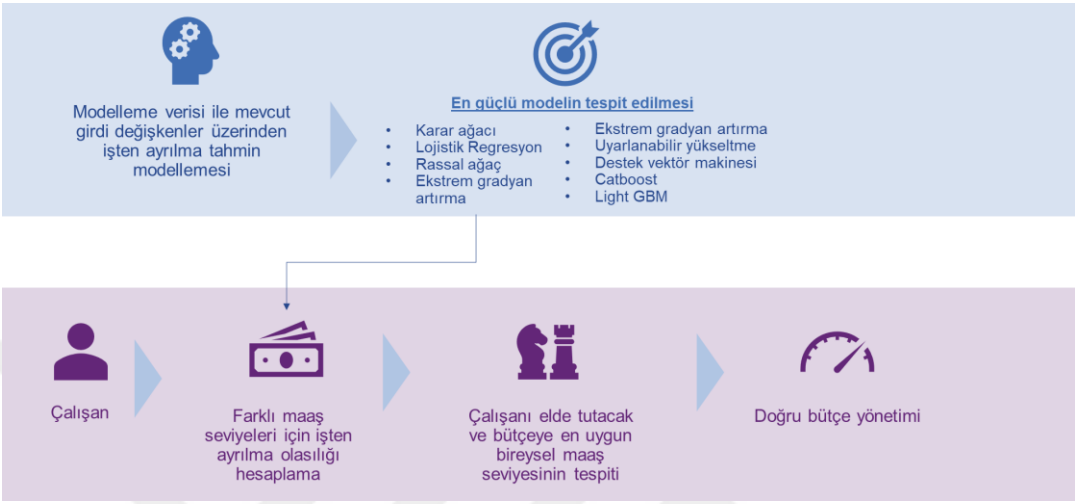
Değişken	Katsayı
Yas	-0.00772
Cinsiyet	0.401339
Medeni_Durum	0.832683
Egitim_Duzeyi	0.112466
Egitim_Alani	0.151618
Departman	0.663638
Is_Rolu	-0.03854
Is_Seviyesi	0.069695
Sirkette_Calisilan_Sure	0.088193
Mevcut_Rolde_Calisilan_Sure	-0.14326
Toplam_Tecrube	-0.07997
Mevcut_Yoneticiyle_Calisilan_Sure	-0.04846
Son_Terfiden_Gecen_Sure	0.142573
Calisilan_Firma_Sayisi	0.206488
Yuzde_Maas_Artisi	-0.01609
Performans_Notu	-0.03743
Egitim_Sayisi	-0.20928
Is_Hayat_Dengesi	-0.14529
Ise_Dahiliyet	-0.39799
Fazla_Mesai	1.870751

Hisse_Seviyesi	-0.04861
Ev_Is_Mesafesi	0.039258
Is_Seyahati	0.030562
Is_Memnuniyeti	-0.2703
Ortam_Memnuniyeti	-0.2726
Iliski_Memnuniyeti	-0.12373
Saatlik_Ucret	-0.00052
Gunluk_Ucret	-0.00014
Aylik_Ucret	6.26E-06
Aylik_Gelir	-2.5E-05
Terfi_Etmis	-0.28286
Rol_Degistirmis	-0.03689
Yonetici_Degistirmis	0
Is_Degistirmis	0.363241
Is_Degistirme_Suresi	-0.08086
Is_Degistirme_Suresine_Yakinlik	-0.0013
Ise_Baslama_Yasi	0.002274
Sirkete_Baslama_Yasi	-0.01272
Ortalama_Memnuniyet	-0.40707
Genc_Olmayan_Bekar	-0.22959
Cok_Genc_Evli	0
Az_Kazanan_Evli	0.601961
Az_Kazanan_Bekar	1.236037
Cok_Kazanan_Bekar	0
Sene_Basina_Egitim_Sayisi	0.24034

5.5. Duyarlılık Analizi

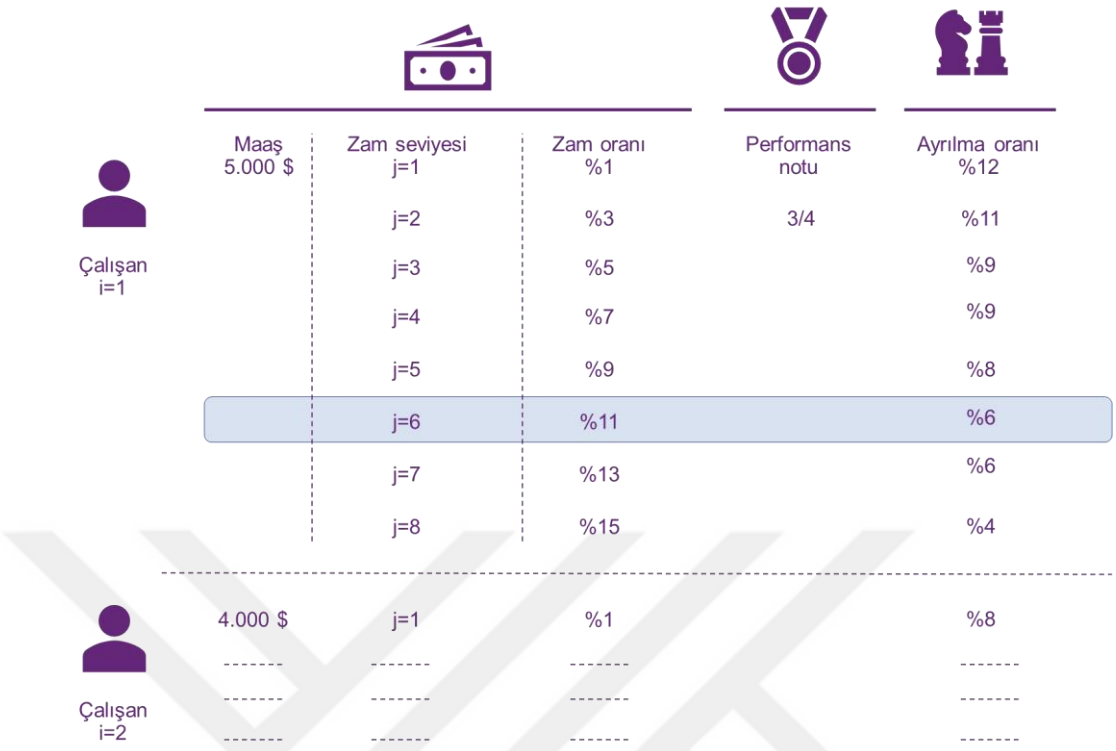
Bu bölümde, model değişkenleri arasında yer alan gelir değişkenleri üzerinde duyarlılık analizi yapılarak bu bilgideki değişimin nihai çıktı olan işten ayrılma durumu üzerindeki etkisi analiz edilecektir. Özellikle maaş değişkeni için farklı alternatif maaş tutarları altında çalışanların işten ayrılma olasılıkları hesaplanacak ve maaşa duyarlı ve kaybedilmemesi gereken yüksek performanslı

çalışanların işten ayrılmasının önüne geçilmesi için bir çalışma yapılacaktır. Bu çalışmanın genel akışı aşağıdaki şekildedir.



Şekil 53. Optimizasyonun Genel Kullanım Akışı

Tek bir müşteri örneği üzerinden akış aşağıdaki şekilde açıklanabilir. i. çalışan için j. maaş seviyesi belirlendiğinde zam oranı karşısında görülen değer olacaktır ve buna göre de yapay zeka modelleri tekrar çalıştırılarak ayrılma oranı tahmin edilecektir. Yapılan tahminler doğrultusunda çalışanın performansını da dikkate alarak en uygun maaş seviyesi belirlenecek ve bütçe altında kalınarak en optimal çalışan maaşları ile şirkete fayda sağlanacaktır.



Şekil 54. Optimizasyonun Genel Kullanım Akışı (2)

Önceki bölümlerde detayları paylaşılan analitik modeller kullanılarak maaş seviyesi %1 ile %15 arasındaki alternatif oranlarda artırıldığındaki işten ayrılma olasılığı tekrar hesaplandığında modellerin hangisinin maaş seviyesine karşı en duyarlı olduğu bulunabilmiştir. Bu analize göre, karar ağacı, ekstrem gradyan artırma ve Light GBM modelleri maaş seviyesinin etkisini en yüksek derece işten ayrılma üzerinde gösteren modeller olarak karşımıza çıkmıştır. Optimizasyon modelinin temellendirileceği model algoritması olarak da Light GBM modelinin seçilmesine karar verilmiştir.

Tablo 30. Algoritma Bazında Maaş Artışına Göre Duyarlılık Analizi

Algoritma	%X aylık gelir artışı işten ayrılma oranını tüm kitlede ne kadar etkiliyor?							
	%1	%3	%5	%7	%9	%11	%13	%15
Karar Ağacı	%1.24	%3.72	%5.37	%6.20	%8.26	%8.26	%8.68	%9.50
Rassal Ağaç	%0.27	%0.38	%0.61	%1.01	%1.25	%1.41	%1.33	%1.33
Lojistik Regresyon	%0.08	%0.24	%0.41	%0.57	%0.73	%0.89	%1.05	%1.21
Ekstrem Gradyan Artırma	%0.84	%1.88	%3.10	%3.67	%4.77	%5.47	%5.49	%6.42
Uyarlanabilir Yükseltme	%0.01	%0.02	%0.03	%0.04	%0.05	%0.08	%0.08	%0.08
Catboost	%0.46	%1.06	%1.70	%2.31	%2.93	%3.43	%3.82	%4.25
Light GBM	%0.65	%1.25	%2.15	%3.08	%4.13	%4.89	%5.27	%5.87

5.6. Optimizasyon

Optimizasyon adımında çalışan değerinin maksimize edilmesi amaçlanmıştır. Bu nedenle, çalışan performans notunu ve işte kalma olasılığını içerisine alan bir fonksiyon yaratılarak maksimize edilmesi planlanmıştır.

Seçili amaç fonksiyonunu maksimize etme problemindeki kısıtlar ise şunlar olmuştur:

- Her çalışanın sadece bir tek maaş artış seviyesi olabilir.
- İlgili maaş artış seviyeleri çalışanların mevcut maaşları üzerine uygulandığında toplam bütçe aşılmamalıdır.
- Yeni bütçe, eski bütçe üzerine sadece enflasyon etkisi eklenmiş şekilde olmalıdır. Ancak, doğrudan enflasyon katsayısı eklemek yerine çeşitli artış oranları denenerek bir duyarlılık analizi yapılacaktır. Böylece enflasyondan daha düşük bir oranda artış ile de çalışanların işten ayrılma durumu kontrol altında tutulabiliyorsa bütçe tasarrufu sağlama yolu da açılmış olacaktır.
- Enflasyon, resmi istatistik verisine göre baz alınmalıdır.
- Maaş artış seviyesi 0-1 şeklinde olmalıdır.
- Enflasyona dayalı maaş artış seviyeleri tamsayı değerler olmalı ve daha önce belirlenmiş 8 değer arasından biri olmalıdır.

Tablo 31. Optimizasyon Modeli Parametreleri

Kısaltma	Açıklama
n	Çalışan (1-N)
m	Maaş seviyesi (1-M)
P_n	n. çalışan için performans notu (3,4)
A_{nm}	n. çalışan için m. maaş seviyesinde işten ayrılma olasılığı (0-1)
D_{nm}	n. çalışan için m. maaş seviyesinde olma durumu (0,1)
Z_m	m maaş seviyesindeki maaş artışı (%1 - %15)
S_{nm}	n. çalışan için m. maaş seviyesi
T	Bütçe
e	Enflasyon oranı

Amaç fonksiyonu:

$$max \sum_{n=1}^{N} \sum_{m=1}^{M} P_n \, x \, (1 - A_{nm}) x \, D_{nm}$$

Kısıtlar:

$$\sum_{m=1}^{M} D_{nm} = 1 \qquad \forall \, n \, için \quad (\mathbf{1})$$

$$\sum_{n=1}^{N} \sum_{m=1}^{M} S_{nm} \, x \, D_{nm} \leq T_{yeni} \quad (\mathbf{2})$$

$$T_{yeni} = T_{eski} \, x \, (1 + e) \quad (\mathbf{3})$$

$$S_{nm} = S_0 \, x \, (1 + Z_m) \quad (\mathbf{4})$$

$$e = \%8{,}6 \quad (\mathbf{5})$$

$$D_{ij} = 0 \; veya \; 1 \quad (\mathbf{6})$$

$$M = 9 \; (\mathbf{7})$$

$$Z_0 = \%0, \qquad Z_1 = \%1, \qquad Z_2 = \%3, \qquad Z_3 = \%5, \qquad Z_4 = \%7,$$

$$Z_5 = \%9, \; Z_6 = \%11, \; Z_7 = \%13, \; Z_8 = \%15 \; (\mathbf{8})$$

Optimizasyon çalışması için Python içerisindeki Pyomo kütüphanesi kullanılmıştır. Bunun için şu adımlar izlenmiştir:

1. Model Python dilinde kodlanmıştır.
2. Çözücü üzerinde model koşturulmuştur.
3. Model sonuçları analiz edilmiştir.

Python Pyomo kütüphanesi içerisinde optimizasyon için şu Python sınıflarının tanımlanması gerektiği için problem formülasyonunda öncelikle bu adım izlenmiştir (Bynum, M. L. vd., 2021):

- Var: Modeldeki optimizasyon değişkenleridir
- Objective: Minimize veya maksimize edilmeye çalışılan değerdir
- Constraint: Modeldeki kısıtlardır
- Set: Modeli tanımlayan veri kümesidir

- Param: Modeli tanımlayan parametre verisidir

Python modelinin ilgili elemanları aşağıdaki kod içerisinde özetlenmiştir:

```
def wage_optimizer_model(N, M, P, A, S):
    model = pyo.ConcreteModel(name="(wage_optimizer_model)")
    model.d = pyo.Var(N, M, within=pyo.Binary)
    def obj_rule(mdl):
        return sum(P[n,m]*(1-A[n,m])*mdl.d[n,m] for n in N for m in M) model.obj =
    pyo.Objective(rule=obj_rule, sense=pyo.maximize)
    def tek_maas_kisiti(mdl, n):
        return sum(mdl.d[n,m] for m in M) == 1
    model.tek_maas_kisiti = pyo.Constraint(N, rule=tek_maas_kisiti)
    def butce_kisiti(mdl):
        return sum(S[n,m]*mdl.d[n,m] for n in N for m in M) <= T_yeni
    model.butce_kisiti = pyo.Constraint(rule=butce_kisiti)
return model
```

Optimizasyon modelinin çıktısı çalışan bazlı zam alıp almayacakları veya zam seviyesi olmaktadır. Buna göre örnek bir model çıktısı aşağıdaki şekildedir:

Tablo 32. Optimizasyon Modeli Sonucu

Satır Etiketleri	Zam yok	%1 zam	%3 zam	%5 zam	%7 zam	%9 zam	%11 zam	%13 zam	%15 zam
Employee_Nr_1022	1	0	0	0	0	0	0	0	0
Employee_Nr_1023	0	0	0	0	0	0	0	0	1
Employee_Nr_1024	0	0	1	0	0	0	0	0	0
Employee_Nr_1025	1	0	0	0	0	0	0	0	0
Employee_Nr_1026	1	0	0	0	0	0	0	0	0
Employee_Nr_1027	1	0	0	0	0	0	0	0	0
Employee_Nr_1028	0	0	0	0	0	0	0	0	1
Employee_Nr_1029	0	0	0	0	0	0	1	0	0
Employee_Nr_103	0	0	0	0	0	0	0	0	1
Employee_Nr_1030	1	0	0	0	0	0	0	0	0
Employee_Nr_1031	1	0	0	0	0	0	0	0	0
Employee_Nr_1032	0	0	0	0	1	0	0	0	0
Employee_Nr_1033	0	0	0	0	0	0	0	0	1

Bu optimizasyon modeli üzerinde şirket bütçesinin farklı alternatifleri için modelin yeniden çalıştırılması sonucunda hem zam yapılan kişi sayısının arttığı hem de kişi bazlı önerilen zam oranlarının da yukarı yönlü değiştiği gözlemlenmiştir. Öncelikle şirketin şu anki maaş bütçesi aylık olarak 9.5M dolar olduğundan 9.6M dolar bütçe ile ilk model çalıştırılmıştır. Bundan sonra ise 9.7M, 9.8M, 9.9M, 10M, 10.1M, 10.2M, 10.3M, 10.4M, 10.5M dolar bütçe kısıtları ile yeniden modelleme yapılmıştır. Her bütçe kısıtı altında da zam almaya hak kazanan çalışan oranları aşağıdaki tabloda verilmiştir.

Tablo 33. Bütçe için Duyarlılık Analizi Sonuçları

Yeni Bütçe	Zam yapılan çalışan oranı
9.6M	10.8%
9.7M	28.6%
9.8M	42.7%
9.9M	49.4%
10M	54.6%
10.1M	58.9%
10.2M	65.8%
10.3M	70.3%
10.4M	73.7%
10.5M	76.7%

6. SONUÇ

Bu çalışmada literatürde son yıllarda odaklanılmaya başlansa da insan kaynakları alanındaki nadir analitik konulardan biri olan çalışan kaybına odaklanılmıştır.

Öncelikle, insan kaynakları literatürü taranmış ve genelden özele doğru giden bir yaklaşım ile çalışan kaybı konusuna odaklanılmaya çalışılmıştır. Literatür taramasında hem Türkiye hem de dünyadaki saygın yayınevleri ve dergiler önceliklendirilmiştir.

Sonraki bölümlerde, veriye dayalı, analitik bir yaklaşım ile çalışanların işten ayrılması tahmin edilmiştir. Piyasada en çok kullanılan makine öğrenmesi algoritmaları kullanılarak Python programlama dili ile modeller geliştirilmiş ve modellerin performans metrikleri karşılaştırılarak en uygun modelleme algoritmasının hangisi olduğu belirlenmiştir. En uygun algoritmanın sonucunda verilen çalışan seviyesindeki işten ayrılma ihtimalleri başta olmak üzere çalışan performansları ve güncel durumdaki maaş bilgileri göz önüne alınarak şirketin toplam maaş bütçesini geçmeyecek şekilde maaş seviyeleri yeniden organize edilmiştir. Çalışma, amaç fonksiyon ve kısıtlar gereği, maaş seviyesine göre hassas olan, performansı yüksek ve işten ayrılma ihtimali yüksek çalışanları önceliklendiren bir yapıda olmuştur. Böylece, performans odaklı bir ücretlendirme politikası izlenerek çalışanların işten ayrılmasının önüne geçilebileceği gösterilmiştir.

Bu çalışma üzerine, ileride yapılacak çalışmalarda daha büyük bir veri kümesi ile çalışılması, çalışana dair daha geçmişe giden tarihsel bir veri kullanılması, o günün şartlarındaki en güncel algoritmaların değerlendirilmesi ek faydalar getirecektir.

7. KAYNAKÇA

Acharya, A., Singh, S. K., Pereira, V., & Singh, P. (2018), "Big data, knowledge co-creation and decision making in fashion n industry", *International Journal of Information Management,* 42: 90–101.

Agrawal, T. (2021), *Hyperparameter Optimization in Machine Learning: Make your machine learning and deep learning models more efficient*. Bangalore, Hindistan: Apress.

Aktan, E. (2018), "Büyük Veri: Uygulama Alanları, Analitiği ve Güvenlik Boyutu", *Ankara Üniversitesi Bilgi Yönetimi Dergisi*: 1–22.

Alao, D., Adeyemo, A. B. (2013), "Analyzing employee attrition using decision tree algorithms", *Computing, Information Systems & Development Informatics,* 4/ 1: 17-28.

Alamsyah, A., Salma, N. (2018), "A Comparative Study of Employee Churn Prediction Model", *4th International Conference on Science and Technology held at Indonesia*: 1-4.

Aliyu, O. A., Nyadzayo, M. W. (2018), "Reducing employee turnover intention: a customer relationship management perspective", *Journal of Strategic Marketing,* 26/3: 241-257.

Al-Shiakli, S. (2019), Big Data Analytics: A Literature Review Perspective, Yüksek Lisans Tezi, Luleå, Luleå University of Technology, Department of Computer Science, Electrical and Space Engineering.

Assunção, M. D., Calheiros, R. N., Bianchi, S., Netto, M. A. S., Buyya, R. (2015), "Big Data computing and clouds: Trends and future directions", *Journal of Parallel and Distributed Computing,* 79/80: 3–15.

Bohdal-Spiegelhoff, U., Barry, L., Erickson, R., Lamoureux, K. (2020, 7 Ekim), "Talent Acquisition Revisited: Deploy New Approaches for the New Battlefield", http://dupress.com/articles/hc-trends-2014-talent-acquisition/ adresinden alındı

Bellon, M. (2016). "The characteristics of worker flows by firm growth: empirical evidence from a matched firm-worker dataset from France", (Columbia University). https://matthieubellon.com /docs/WorkerFlowsFirmGrowth.pdf

Bhuva, K., Srivastava, K. (2018), "Comparative Study of the Machine Learning Techniques for Predicting the Employee Attrition", *International Journal of Research and Analytical Reviews,* 5/3: 568-577.

Boyd, D., & Crawford, K. (2012), "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon", *Information Communication and Society*, 15/5: 662–679.

Böckerman, P. & Ilmakunnas, P. (2009), "Job disamenities, job satisfaction, quit intensions and actual separations: putting the pieces together", *Industrial Relations*, 48: 73-96.

Brownlee, J. (2020), *Data Preparation for Machine Learning: Data Cleaning, feature Selection, and Data Transforms in Python*, Machine Learning Mystery.

Bryson, S., vd. (1999), "Visually Exploring Gigabyte Data Sets in Real Time", *Communications of the ACM*, 42/8: 82-90.

Burgess, S., Lane, J. & Stevens, D. (2000), "Job Flows, Worker Flows and Churning", *Journal of Labor Economics*, 18: 473-502.

Burgess, S., Lane, J. & Stevens, D. (2001), "Churning dynamics: an analysis of hires and separations at the employer level", *Labour Economics*, 8: 1-14.

Bynum, M. L., Hackebeill, G. A., Hart, W. E., Laird, C. D., Nicholson, B. L., Siirolaa, John. D., Watson, J.-P., & Woodruff, D. L. (2021), *Pyomo - Optimization Modeling in Python*, 67, Cham, Switzerland: Springer Optimization and Its Applications.

CIPD (2020), *Resourcing and talent planning survey 2020*, Chartered Institute of Personnel and Development: London.

CIPD (2020, 10 Eylül), "People analytics: Understand what people analytics is, different types of HR data, and the importance of people analytics strategy and process", https://www.cipd.co.uk/knowledge/strategy/analytics/factsheet adresinden alındı.

Clark, A. E. (2001), "What really matters in a job? Hedonic measurement using quit data", *Labour Economics*, 8: 223-242

Corbett, P., Schroeck, M., & Shockley, R. (2021, 13 Ocak). "Analytics: The real-world use of big data in insurance - How innovative insurance organizations extract value from uncertain data", www.sbs.ox.ac.uk adresinden alındı

Cottini, E., Kato, T. & Westergaard-Nielsen, N. (2011), "Adverse workplace conditions, high-involvement work practices and labor turnover: evidence from Danish linked employer–employee data", *Labour Economics*, 18: 872–880.

Cox, M., Ellsworth, D. (1997), "Application-Controlled Demand Paging for Out-of-Core Visualization", *In Proceedings of Visualization'97*, IEEE Press: 235-244.

Dale-Olsen, H. (2016), Employee Turnover. *Research Handbook of Employee Turnover*, Cheltenham, UK: Edwar Elgar Publishing.

Dale-Olsen, H. (2006), "Wages, fringe benefits and worker turnover", *Labour Economics*, 13: 87-105.

Davis, S. J. & Haltiwanger, J. (1999), "Gross job flows", Ashenfelter, O., Card, D. (Ed.), *Handbook of Labor Economics*, 3B, Amsterdam, Netherlands and New York, USA: Elsevier: 2711-2805.

Davis, S.J. & J. Haltiwanger (2014), "Labor market fluidity and economic performance", NBER Working paper #20479.

Deeba, F. (2020), "The Role of Data Analytics in Talent Acquisition and Retention with Special Reference to SMEs in India: A Conceptual Study", *International Conference on "Management Imperatives for Sustainable Growth – Transformation Through Technology held at ICFAI Business School on August 24, 2019*, Gurgaon, Haryana, India.

Delfgaauw, J. (2007), "Where to go? Workers' reasons to quit and intra- vs. interindustry job mobility", *Applied Economics*, 39: 2057-2067.

Fisher, D., DeLine, R., Czerwinski, M., & Drucker, S. (2012), "Interactions with Big Data Analytics", Interactions.

Fitz-enz, J. (2012), "Why They Leave: What the Research Reveals", Branham, L. (Ed.), *7 Hidden Reasons Employees Leave: How to Recognize the Subtle Signs and Act Before It's Too Late*, New York: Amacom: 21-36.

Fitz-enz, J., Mattox, J. R. II. (2014), *Predictive Analytics for Human Resources*. New Jersey, USA: John Wiley & Sons.

Fletcher, L., Alfes, K., Robinson, D. (2018), "The relationship between perceived training and development and employee retention: the mediating role of work attitudes", *The International Journal of Human Resource Management*, 29/18: 2701-2728.

Gahi, Y., Guennoun, M., & Mouftah, H. T. (2016), "Big Data Analytics: Security and privacy challenges", *Proceedings - IEEE Symposium on Computers and Communications*, 2016-August: 952–957.

Gandomi, A., Haider, M. (2015), "Beyond the hype: Big data concepts, methods, and analytics", *International Journal of Information Management*, 35: 137-144.

Gee, S. (2015), *Fraud and Fraud Detection: A Data Analytics Approach* (1. Baskı), John Wiley & Sons.

Geeksforgeeks, (2022, 12 Ocak), https://www.geeksforgeeks.org/difference-between-data-mining-and-data-analysis/ adresinden alındı.

Gene, P. (2015), *Optimize Your Greatest Asset - Your People: How to Apply Analytics to Big Data to Improve Your Human Capital Investments*. New Jersey: John Wiley & Sons.

Gosney, M. W., Hughes, C. (2016), *The History of Human Resource Development: Understanding the Unexplored Philosophies, Theories, and Methodologies,* London: Palgrave Macmillan.

Govekar, P., L. & Govekar, M. A. (2012), "The Parable of the Pig Iron: Using Taylor's Story to Teach the Principles of Scientific Management", *Journal of Higher Education Theory and Practice,* 12: 73-83.

Güçlükol, S. (2020), Forecasting Customer Service Demand by Machine Learning with Real Life Application, Yüksek Lisans Tezi, İstanbul, Sabancı Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi.

Gürkan, S., Dazlak, B. D. (2019), "İnsan Kaynakları Temin ve Seçiminde Analitik Hiyerarşi Prosesinin Kullanılması ve Oteller Üzerine Bir Uygulama", *Selçuk Üniversitesi Sosyal Bilimler Meslek Yüksek Okulu Dergisi*, 22/ 2: 774-783.

Harvard Business Review Analytic Services (2014), HR *Joins the Analytics Revolution*, Harvard Business School Publishing: Boston.

Hazırbaba, N. (2020), Organizasyonun Büyük Veri Analitiği Yeteneği Doğrultusunda Tasarımı: Bir Havayolu Örneği, Doktora Tezi, İstanbul, İstanbul Ticaret Üniversitesi, Sosyal Bilimler Enstitüsü.

He, Y., Yu, F. R., Zhao, N., Yin, H., Yao, H., & Qiu, R. C. (2016), "Big Data Analytics in Mobile Cellular Networks", *IEEE Access*, 4: 1985–1996.

Hendrick, R. Z. ve Raspiller, E. E. (2011), "Predicting Employee Retention Through Preemployment Assessment. *Community College Journal of Research and Practice*", 35/11: 895-908.

Igual, L., Seguí, S. (2017), *Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications*, Cham, Switzerland: Springer.

Isson, J. P., Harriott, J. S. (2016), *People Analytics in the Era of Big Data*, New Jersey, USA: John Wiley & Sons.

Işık, H. (2020), Sales Forecasting in Fashion Retail Industry with Classical and Machine Learning Methods, Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi, Sosyal Bilimler Enstitüsü.

Jafari, R. (2022), *Hands-On Data Preprocessing in Python: Learn How to Effectively Prepare Data for Successful Data Analytics*, Birmingham, UK: Packt Publishing.

Jantan, H., Hamdan, A. R. & Othman, Z. A. (2011), "Towards applying Data Mining Techniques for Talent Management", *Paper presented at International Conference on Computer Engineering and Applications.*

Jiang, H. (2021), *Machine Learning Fundamentals: A Concise Introduction*, Cambridge, UK: Cambridge University Press.

Jung, A. (2022), *Machine Learning The Basics: Foundations, Methodologies and Applications*, Singapore: Springer.

Kalhori, M., Kargar, M. J. (2017), "An Analytical Employee Performance Evaluation Approach in Office Automation and Information Systems", Tavana, M., Szabat, K., Puranam, K. (Ed.), *Organizational Productivity and Performance Measurements Using Predictive Modeling and Analytics*, Hershey PA, USA: IGI Global: 324-343.

Karaboğa, T. (2020), Büyük Veri Analitiği Yönetsel Kabiliyetlerinin Firma Performansına Etkisi: Veri Odaklı Kültür ve Büyük Veri Strateji Uyumunun Aracılık Etkisi, Doktora Tezi, İstanbul, Yıldız Teknik Üniversitesi, Sosyal Bilimler Enstitüsü.

Katsieka, E., Theodosiou, M. & Morgan, R., E. (2015), "Why people quit: Explaining employee turnover intentions among export sales managers", *International Business Review*, 24: 367-379.

Kemp, R. (2014), "Legal aspects of managing Big Data", *Computer Law and Security Review*, 30/5: 482–491.

Khan, S., Liu, X., Shakil, K. A., & Alam, M. (2017), "A survey on scholarly data: From big data perspective", *Information Processing and Management*, 53/4: 923–944.

Kısaoğlu, Z. Ö. (2014), Employee Turnover Prediction Using Machine Learning Based Methods, Yüksek Lisans Tezi, Ankara, Ortadoğu Teknik Üniversitesi.

Kihlstrom, J., F. & Cantor, N. (2000), "Social Intelligence", *The Psychometric View*.

Kishore, K., Majumdar, M, & Kiran, V. (2012), "Innovative HR Strategies for SMEs", *IOSR Journal of Business and Management*, 2/ 6: 01-08.

Kozak, J. (2019), *Decision Tree and Ensemble Learning Based on Ant Colony Optimization (781)*, Cham: Springer International Publishing.

Kubick, W. R. (2012), "Big Data, Information and Meaning", *Applied Clinical Trials*, 21/2: 26–28.

Kumar, A., & Jain, M. (2020), *Ensemble Learning for AI Developers: Learn Bagging, Stacking and Boosting Methods with Use Cases. In Ensemble Learning for AI Developers*, Gurugram, Hindistan: Apress.

Laney, D. (2021, 14 Mart), "3-D data management: Controlling data volume, velocityand variety. Application Delivery Strategies by META Group Inc", http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf adresinden alındı.

Latifian, A. (2022), "How does cloud computing help businesses to manage big data issues". *Kybernetes*.

Lee, T. W., vd. (2018), "Managing employee retention and turnover with 21st century ideas", *Organisational Dynamics*, 47: 88-98.

Levy-Garboua, L., Montmarquette, C. & Simonnet, V. (2007), "Job satisfaction and quits", *Labour Economics*, 14: 251-268.

Madaan, N., Kumar, U., & Jha, S. K. (2020), "Big Data Analytics: A Literature Review Paper", *International Journal of Engineering Research & Technology (IJERT)*: 11–19.

Masood, A. (2021), *Automated Machine Learning: Hyperparameter optimization, neural architecture search, and algorithm selection with cloud platforms*, Birmingham: Packt Publishing Limited.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). "Big data: The next frontier for innovation, competition, and productivity", McKinsey.

McNulty, K. (2021), *Handbook of Regression Modeling in People Analytics With Examples in R and Python*, Boca Raton, USA: CRC Press.

Medium (2011, 12 Ocak), "Design Thinking Humanizes Data Science More", https://medium.com/technicity/design-thinking-humanizes-data-science-more-5a666119c8b1 adresinden alındı.

Mikalef, P., Boura, M., Lekakos, G., & Krogstie, J. (2019), “Big data analytics and firm performance: Findings from a mixed-method approach”, *Journal of Business Research*, 98: 261–276.

Necula, S. C. & Strîmbei, C. (2019), “People Analytics of Semantic Web Human Resource Résumés for Sustainable Talent Acquisition”, *Sustainability*, 11: 3520-3538.

Niu, C., & Wang, L. (2022), “Big data-driven scheduling optimization algorithm for Cyber–Physical Systems based on a cloud platform”, *Computer Communications*, 181: 173–181.

OECD (2009), *How Do Industry, Firm and Worker Characteristics Shape Job and Worker Flows?, OECD Employment Outlook 2009, Tackling the Job Crisis. Paris: Organisation for Economic Co-Operation and Development*, 117–163.

OECD (2010), Institutional and policy determinants of labour market flows, OECD Employment Outlook 2010, Moving Beyond the Jobs Crisis. Paris: Organisation for Economic Co-Operation and Development, 167–210.

Pérez-Campdesuñer, R., De-Miguel-Guzmán, M., Sánchez-Rodríguez, A., García-Vidal, G., Martínez-Vivar, R. (2018), “Exploring neural networks in the analysis of variables that affect the employee turnover in the organization”, *International Journal of Engineering Business Management*, 10: 1-11.

Presbitero, A., Roxas, B., & Chadee, D. (2016), “Looking beyond HRM practices in enhancing employee retention in BPOs: focus on employee–organisation value fit”, *The International Journal of Human Resource Management*, 27/6: 635-652.

Punnoose, R., Ajit, P. (2016), “Prediction of Employee Turnover in Organizations using Machine Learning Algorithms: A case for Extreme Gradient Boosting”, *International Journal of Advanced Research in Artificial Intelligence*, 5/9.

Quinto, B. (2020), “Next-generation machine learning with spark: Covers XGBoost, LightGBM, Spark NLP, distributed deep learning with keras, and more”, *Next-Generation Machine Learning with Spark: Covers XGBoost, LightGBM, Spark NLP, Distributed Deep Learning with Keras, and More*. Carson, USA: Apress Media LLC.

Reilly P. (1996), “Human Resource Planning: An Introduction”, (Institute for Employment Studies Report Report, 312).

Rosen, M. & Spaulding, T. (2009), "Best Practices for Wellness Programs”, *Occupational Health & Safety Online*.

Rothman, D. (2020), *Hands-On Explainable AI (XAI) with Python: Interpret, visualize, explain and interpret reliable AI for fair, secure, and trustworthy AI apps*, Birmingham: Packt Publishing.

Runkler, T. A. (2016), *Data Analytics: Models and Algorithms for Intelligent Data Analysis (2. Baskı)*, München, Germany: Springer.

Sabbagha, M. D. S., Ledimo, O. & Martins, N. (2018), "Predicting staff retention from employee motivation and job satisfaction", *Journal of Psychology in Africa*, 28/ 2: 136-140.

Salunkhe, T. P. (2018). Improving Employee Retention by Predicting Employee Attrition using Machine Learning Techniques, Yüksek Lisans Tezi, Dublin, Dublin Business School, Computing.

Saradhi, V., V. & Palshikar, G., K. (2011), "Employee churn prediction", *Expert Systems with Applications*, 38: 1999-2006.

Schweyer, A. (2004), *Talent Management Systems: Best Practices in Technology Solutions for Recruitment, Retention and Workforce Planning*, Canada: John Wiley & Sons.

Shah, N., Irani, Z. & Sharif A. (2017), "Big data in an HR context: Exploring organizational change readiness, employee attitudes and behaviors", *Journal of Business Research*, 70: 366-378.

Shin, D. H. (2016), "Demystifying big data: Anatomy of big data developmental process", *Telecommunications Policy*, 40/9: 837–854.

Sinclair A. (2004), "Workforce Planning: A literature review", *Institute for Employment Studies Report.*

Singh, P. (2021), *Deploy Machine Learning Models to Production: With Flask, Streamlit, Docker and Kubernetes on Google Cloud Platform*, Bangalore, India: Apress.

Stamolampros, P., vd. (2019), "Job satisfaction and employee turnover determinants in high contact services: Insights from Employees' Online reviews", *Tourism Management*, 75: 130-147.

Statista (2022, 24 Şubat), https://www.statista.com/statistics/871513/worldwide-data-created/#:~:textThe%20total%20amount%20of%20data,replicated%20reached%20a%20new%20high, adresinden alındı

Strohmeier, S. & Piazza, F. (2015), *Human Resource Intelligence und Analytics: Grundlagen, Anbieter, Erfahrungen und Trends*, Wiesbaden: Springer Gabler.

Subrahmanyam, A. (2019), "Big data in finance: Evidence and challenges", *Borsa Istanbul Review* 19/4: 283–287

Tanyaş, M. (2000), *Endüstri Mühendisliğine Giriş (1. Baskı),* İstanbul: İrfan Yayımcılık.

Tyagi, S. (2003), "Using data analytics for greater profits", *Journal of Business Strategy*, 24/3, 12 – 14.

Urwin, P. & Parry, E. (2016), "Two decades of employee retention, tenure and turnover", Sarikadis, G., & Cooper, C. (Ed.), *Research Handbook of Employee Turnover*, Cheltenham, UK: Edwar Elgar Publishing: 59-78.

Visier (2020), *The Age of People Analytics: Survey on Characteristics, Value Achieved, and Leading Practices of Advanced Organizations*, Visier Surveys.

Walford-Wright, G. & Scott-Jackson, W. (2018), "Talent Rising; people analytics and technology driving talent acquisition strategy", *Strategic HR Review*, 17/5: 226-233.

Wu, J. (2012), *Advances in K-means Clustering: A Data Mining Thinking*: Recognizing Outstanding Ph.D. Research, Springer Theses: Berlin.

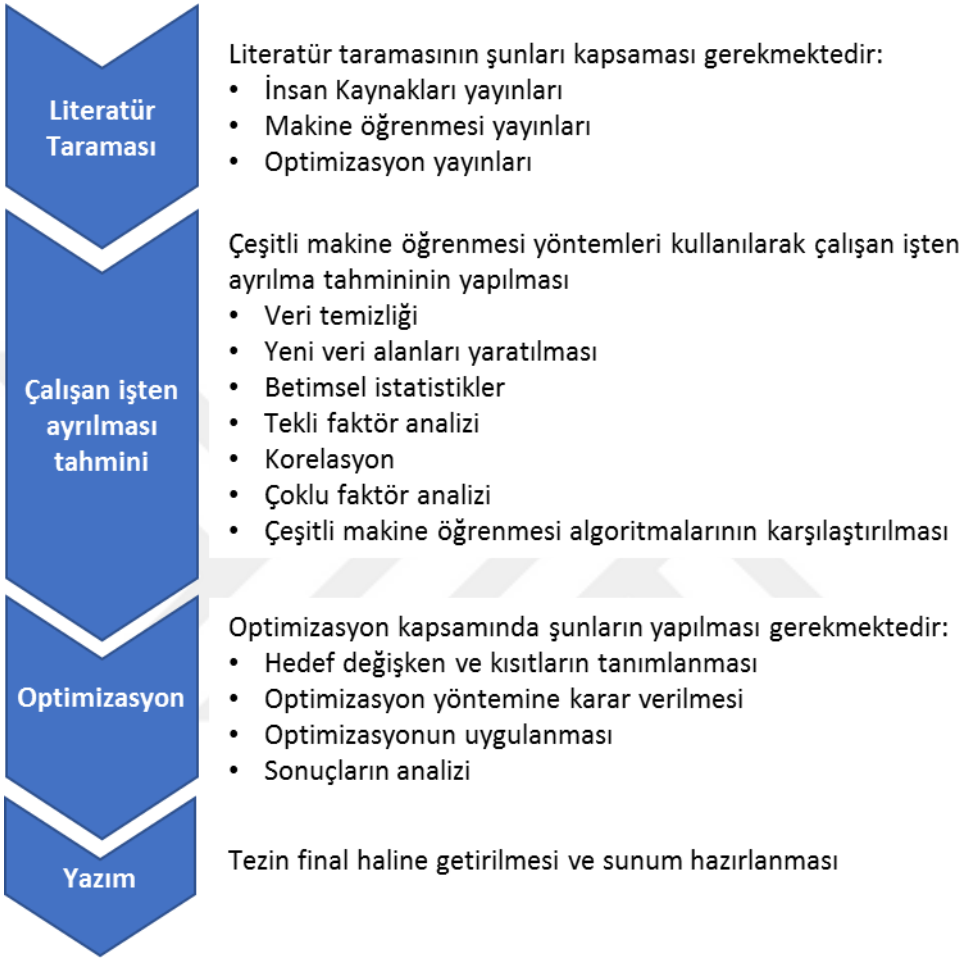
Yang, Y., Li, Y., Chen, R., Zheng, J., Cai, Y., & Fortino, G. (2021), "Risk Prediction of Renal Failure for Chronic Disease Population Based on Electronic Health Record Big Data", *Big Data Research*, 25.

Yang, J. T., Wan, C. S. & Fu, Y. J. (2012), "Qualitative examination of employee turnover and retention strategies in international tourist hotels in Taiwan", *International Journal of Hospitality Management*, 31: 837-848.

Yilmaz, G. (2019), Hilenin Önlenmesi ve Tespitinde Büyük Veri Analitiğinin Kullanımı ve Türkiye'deki Çalışan Hileleri Üzerine Bir Araştırma, Doktora Tezi, İstanbul, Marmara Üniversitesi Sosyal Bilimler Enstitüsü

8. EKLER

Ek 1: Çalışma Planı



Şekil 55. Çalışma Planı

Ek 2: Kodlar

```
import pandas as pd

import numpy as np


path = 'C:/Users/Asus/Desktop/KadirYildiz/Thesis'

path_input = 'C:/Users/Asus/Desktop/KadirYildiz/Thesis/Data'
```

```
path_output = 'C:/Users/Asus/Desktop/KadirYildiz/Thesis/Output'


#read the input dataset
input_data = pd.read_csv(path_input + "/" + "WA_Fn-UseC_-HR-Employee-Attrition.csv")


#Data Preview
#review the shape of the input dataset
print("Verinin boyutları: {}".format(input_data.shape))


#review the columns in the input dataset
input_data.columns


#translation of the variables
input_data.rename(columns={'Age': 'Yas',
                           'Attrition':'Isten_Ayrilma',
                           'BusinessTravel':'Is_Seyahati',
                           'DailyRate':'Gunluk_Ucret',
                           'Department':'Departman',
                           'DistanceFromHome':'Ev_Is_Mesafesi',
                           'Education':'Egitim_Duzeyi',
                           'EducationField':'Egitim_Alani',
                           'EmployeeCount':'Calisan_Sayisi',
                           'EmployeeNumber':'Calisan_No',
                           'EnvironmentSatisfaction':'Ortam_Memnuniyeti',
                           'Gender':'Cinsiyet',
```

```
                    'HourlyRate':'Saatlik_Ucret',
                    'JobInvolvement':'Ise_Dahiliyet',
                    'JobLevel':'Is_Seviyesi',
                    'JobRole':'Is_Rolu',
                    'JobSatisfaction':'Is_Memnuniyeti',
                    'MaritalStatus':'Medeni_Durum',
                    'MonthlyIncome':'Aylik_Gelir',
                    ''MonthlyRate'':'Aylik_Ucret',
                    'NumCompaniesWorked':'Calisilan_Firma_Sayisi',
                    'Over18':'Yas_18_Ustu',
                    'OverTime':'Fazla_Mesai',
                    'PercentSalaryHike':'Yuzde_Maas_Artisi',
                    'PerformanceRating':'Performans_Notu',
                    'RelationshipSatisfaction':'Iliski_Memnuniyeti',
                    'StandardHours':'Standart_Saatler',
                    'StockOptionLevel':'Hisse_Seviyesi',
                    'TotalWorkingYears':'Toplam_Tecrube',
                    'TrainingTimesLastYear':'Egitim_Sayisi',
                    'WorkLifeBalance':'Is_Hayat_Dengesi',
                    'YearsAtCompany':'Sirkette_Calisilan_Sure',
                    'YearsInCurrentRole':'Mevcut_Rolde_Calisilan_Sure',
                    'YearsSinceLastPromotion':'Son_Terfiden_Gecen_Sure',
                    'YearsWithCurrManager':'Mevcut_Yoneticiyle_Calisilan_Sure'},
inplace=True)
```

```
#changing the order of the variables

input_data = input_data[['Calisan_No', 'Yas', 'Yas_18_Ustu' ,'Cinsiyet', 'Medeni_Durum', 'Egitim_Duzeyi', 'Egitim_Alani','Calisan_Sayisi', 'Departman', 'Is_Rolu', 'Is_Seviyesi', 'Sirkette_Calisilan_Sure', 'Mevcut_Rolde_Calisilan_Sure', 'Toplam_Tecrube', 'Mevcut_Yoneticiyle_Calisilan_Sure', 'Son_Terfiden_Gecen_Sure', 'Calisilan_Firma_Sayisi',

'Yuzde_Maas_Artisi','Performans_Notu','Egitim_Sayisi','Is_Hayat_Dengesi','Ise_Dahiliyet', 'Standart_Saatler','Fazla_Mesai','Hisse_Seviyesi','Ev_Is_Mesafesi','Is_Seyahati', 'Is_Memnuniyeti','Ortam_Memnuniyeti','Iliski_Memnuniyeti','Saatlik_Ucret','Gunluk_Ucret','Aylik_Ucret','Aylik_Gelir','Isten_Ayrilma']]


input_data.dtypes

input_data.info()


class color:
   PURPLE = '\033[95m'
   CYAN = '\033[96m'
   DARKCYAN = '\033[36m'
   BLUE = '\033[94m'
   GREEN = '\033[92m'
   YELLOW = '\033[93m'
   RED = '\033[91m'
   BOLD = '\033[1m'
   UNDERLINE = '\033[4m'
   END = '\033[0m'


#Sayısal değişkenlerin dağılımı ve temel istatistikleri
```

```
input_data_numeric_columns = input_data.select_dtypes(include=np.number).columns.tolist()

num_col = []
for column in input_data_numeric_columns:
    if input_data[column].nunique() < 25:
        print("")
        print(color.BLUE + color.BOLD + "Sayısal değişken dağılımı: " + column + color.END)
        print(f"{column} : {input_data[column].unique()}")
        print(f"{column} : Ortalama: {input_data[column].mean()}, Minimum: {input_data[column].min()}, Maksimum: {input_data[column].max()}")
        num_col.append(column)
    elif input_data[column].nunique() >= 25:
        print("")
        print(color.BLUE + color.BOLD + "Sayısal değişken dağılımı: " + column + color.END)
        print(f"{column} : Ortalama: {input_data[column].mean()}, Minimum: {input_data[column].min()}, Maksimum: {input_data[column].max()}")
        num_col.append(column)
    else:
        print("")

from sklearn.preprocessing import LabelEncoder

label = LabelEncoder()
input_data["Isten_Ayrilma"] = label.fit_transform(input_data.Isten_Ayrilma)
```

```
#Sayısal değişkenlerin dağılım grafikleri

input_data_num = input_data[num_col]

import matplotlib
from matplotlib import pyplot
import matplotlib.pyplot as plt
import matplotlib.pylab as pylab

input_data_num.hist(figsize=(25,20))
plt.show()

for column in input_data.columns:
    print(f"{column}: Tekil değer sayısı {input_data[column].nunique()}")
    print("=====================================================")

input_data.drop(['Calisan_Sayisi', 'Calisan_No', 'Yas_18_Ustu', 'Standart_Saatler'], axis="columns", inplace=True)

#Kategorik değişkenlerin dağılımı

object_col = []
for column in input_data.columns:
    if input_data[column].dtype == object and len(input_data[column].unique()) <= 25:
```

```
            print("")
            print(color.BLUE + color.BOLD + "Kategorik değişken dağılımı: " + column + color.END)
            object_col.append(column)
            print(f"{column} : {input_data[column].unique()}")
            print(input_data[column].value_counts())
        object_col.remove('Isten_Ayrilma')

        import seaborn as sns
        sns.set_style("whitegrid")
        plt.style.use("fivethirtyeight")

        #Kategorik değişkenlerin dağılımı -2

        input_data_object = input_data[object_col]

        fig, axes = plt.subplots(round(len(input_data_object.columns) / 4), 4, figsize=(20, 16))

        for i, ax in enumerate(fig.axes):
           if i < len(input_data_object.columns):
            ax.set_xticklabels(ax.xaxis.get_majorticklabels(), rotation=60)
            sns.countplot(x=input_data_object.columns[i], alpha=0.9, data=input_data_object, ax=ax, palette="Blues")

        fig.tight_layout()
```

```
#Korelasyon Matrisi


plt.figure(figsize=(30, 30))
sns.heatmap(input_data.corr(), annot=True, cmap="RdYlGn", annot_kws={"size":15})


##########################################################################
#####
#Hedef Değişken Analizi


from scipy import stats
from scipy.stats import norm, skew
import seaborn as sns
from matplotlib.ticker import PercentFormatter
import matplotlib.ticker as ticker


# İşten ayrılma oranı
labels = input_data.Isten_Ayrilma.value_counts().index
colors = ['blue','deeppink']
explode = [0,0]
sizes = input_data.Isten_Ayrilma.value_counts().values


# pie chart
plt.figure(figsize = (5,5))
plt.pie(sizes, explode=explode, labels=labels, pctdistance=0.8, startangle=90, colors=colors,
autopct='%1.1f%%', textprops={'color':"w",'fontsize': 14})
plt.title('İşten ayrılma oranı', color = 'blue',fontsize = 15)
```

```
fig, ax = plt.subplots()

ax.barh(labels, sizes, align='center', alpha=0.7, color=colors)
ax.set_title('İşten ayrılma oranı', color = 'blue',fontsize = 15)

total = len(input_data['Isten_Ayrilma'])
for p in ax.patches:
    percentage = '{:.1f}%'.format(100 * p.get_width()/total)
    x = p.get_x() + p.get_width() + 0.02
    y = p.get_y() + p.get_height()/2
    ax.annotate(percentage, (x, y))

# Save the figure and show
plt.show()



#Sayısal değişkenler ile hedef değişken ilişkisinin analizi

num_col_to_draw = []
for column in input_data.columns:
   if input_data[column].dtypes != object:
       num_col_to_draw.append(column)

num_col_to_draw.remove("Isten_Ayrilma")
```

```
object_col_to_draw = []
for column in input_data.columns:
    if input_data[column].dtypes == object:
        object_col_to_draw.append(column)

plt.figure(figsize=(20, 20))

for i, column in enumerate(num_col_to_draw, 1):
    plt.subplot(8, 4, i)
    input_data[input_data["Isten_Ayrilma"] == 0][column].hist(bins=35, color='blue', label='Isten_Ayrilma = HAYIR', alpha=0.9)
    input_data[input_data["Isten_Ayrilma"] == 1][column].hist(bins=35, color='deeppink', label='Isten_Ayrilma = EVET', alpha=0.9)
    plt.legend()
    plt.xlabel(column)

plt.figure(figsize=(20, 40))

for i, column in enumerate(object_col_to_draw, 1):
    plt.subplot(8, 4, i)
    input_data[input_data["Isten_Ayrilma"] == 0][column].hist(bins=35, color='blue', label='Isten_Ayrilma = HAYIR', alpha=0.9)
    input_data[input_data["Isten_Ayrilma"] == 1][column].hist(bins=35, color='deeppink', label='Isten_Ayrilma = EVET', alpha=0.9)
    plt.legend()
```

```
        plt.xlabel(column)
        plt.xticks(rotation=45)

    #Hedef değişken ile girdi değişkenlerin korelasyon analizi
    input_data.drop('Isten_Ayrilma',
axis=1).corrwith(input_data.Isten_Ayrilma).plot(kind='barh', figsize=(10, 7))

    #İşten Ayrılma Durumuna Göre Sayısal Değişkenler için Dağılım Kontrolü

    for i, column in enumerate(num_col_to_draw, 1):
        (mu, sigma) = norm.fit(input_data.loc[input_data['Isten_Ayrilma'] == 0, column])
        print(column + ' İşten Ayrılma = HAYIR: Ortalama = {:.1f} ve Standart Sapma =
{:.1f}'.format(mu, sigma))
        (mu, sigma) = norm.fit(input_data.loc[input_data['Isten_Ayrilma'] == 1, column])
        print(column + ' İşten Ayrılma = EVET: Ortalama = {:.1f} ve Standart Sapma =
{:.1f}'.format(mu, sigma))

    #İşten ayrılma durumu bazında değişkenlerin dağılımı

    plt.figure(figsize=(20, 40))

    for i, column in enumerate(num_col_to_draw, 1):
        plt.subplot(15, 2, i)
        plt.style.use('seaborn-white')
        plt.grid(True, linestyle="-", color='grey', alpha=0.7)
```

```
            sns.kdeplot(input_data.loc[input_data['Isten_Ayrilma'] == 0, column], color='blue', label='İşten Ayrılma = HAYIR', alpha=0.7)
            sns.kdeplot(input_data.loc[input_data['Isten_Ayrilma'] == 1, column], color='deeppink', label='İşten Ayrılma = EVET', alpha=0.7)
            plt.rcParams['lines.linewidth'] = 2
            plt.xlabel(column)
            plt.ylabel('Çalışan Sayısı')
            plt.title(column + ' İşten ayrılma durumu bazında değişkenlerin dağılımı')
            plt.legend(loc='upper right', fancybox=True, framealpha=1, shadow=False, borderpad=1)
            plt.tight_layout();


        #Distribution of attritions for each field

        plt.figure(figsize=(20, 40))

        for i, column in enumerate(object_col_to_draw, 1):
            plt.subplot(12, 4, i)
            temp = input_data[[column]]
            temp['Ayrılanların_Sayısı'] = input_data[input_data['Isten_Ayrilma'] == 1].shape[0]
            temp_ratio = temp['Ayrılanların_Oranı'] = temp[column].value_counts() / temp['Ayrılanların_Sayısı']
            temp_2 = pd.DataFrame(columns=["Alan", "% Ayrılma"])

            j=0
```

```
            for field in list(input_data[column].unique()):
                ratio = input_data[(input_data[column]==field)&(input_data['Isten_Ayrilma']==1)].shape[0] / input_data[input_data[column]==field].shape[0]
                temp_2.loc[j] = (field, ratio*100)
                j += 1
                print("{} için, işten ayrılanların oranı {:.2f}%".format(field, ratio*100))
                temp_3 = temp_2.groupby(by="Alan").sum()
                temp_3.plot(kind='bar',title=column)
                plt.xticks(rotation=45)

        ##########################################################################
#####
        #Değişken zenginleştirme

        input_data_2=input_data.copy()

        # Yeni değişken türetme -1

        def Terfi_Etmis(data) :
            if data['Son_Terfiden_Gecen_Sure'] > 0:
                return 1
            else:
                return 0
        input_data_2['Terfi_Etmis'] = input_data_2.apply(lambda input_data_2:Terfi_Etmis(input_data_2) ,axis = 1)
```

```
# Yeni değişken türetme -2

def Rol_Degistirmis(data) :
    if data['Mevcut_Rolde_Calisilan_Sure'] < data['Sirkette_Calisilan_Sure']:
        return 1
    else:
        return 0
input_data_2['Rol_Degistirmis'] = input_data_2.apply(lambda input_data_2:Rol_Degistirmis(input_data_2) ,axis = 1)

# Yeni değişken türetme -3

def Yonetici_Degistirmis(data) :
    if data['Mevcut_Yoneticiyle_Calisilan_Sure'] < data['Sirkette_Calisilan_Sure']:
        return 1
    else:
        return 0
input_data_2['Yonetici_Degistirmis'] = input_data_2.apply(lambda input_data_2:Yonetici_Degistirmis(input_data_2) ,axis = 1)

# Yeni değişken türetme -4

def Is_Degistirmis(data) :
    if data['Sirkette_Calisilan_Sure'] < data['Toplam_Tecrube']:
        return 1
```

```
    else:
        return 0
input_data_2['Is_Degistirmis'] = input_data_2.apply(lambda input_data_2:Is_Degistirmis(input_data_2) ,axis = 1)

# Yeni değişken türetme -5

def Is_Degistirme_Suresi(data) :
    if data['Calisilan_Firma_Sayisi'] >0:
        return (data['Toplam_Tecrube'] - data['Sirkette_Calisilan_Sure'])/(data['Calisilan_Firma_Sayisi'])
    else:
        return 0
input_data_2['Is_Degistirme_Suresi'] = input_data_2.apply(lambda input_data_2:Is_Degistirme_Suresi(input_data_2) ,axis = 1)

# Yeni değişken türetme -6

def Is_Degistirme_Suresine_Yakinlik(data) :
    if data['Is_Degistirme_Suresi'] >0:
        return data['Sirkette_Calisilan_Sure']/data['Is_Degistirme_Suresi']
    else:
        return 0
input_data_2['Is_Degistirme_Suresine_Yakinlik'] = input_data_2.apply(lambda input_data_2:Is_Degistirme_Suresine_Yakinlik(input_data_2) ,axis = 1)
```

```
# Yeni değişken türetme -7

input_data_2['Ise_Baslama_Yasi'] = (input_data_2['Yas'] - input_data_2['Toplam_Tecrube'])

# Yeni değişken türetme -8

input_data_2['Sirkete_Baslama_Yasi'] = (input_data_2['Yas'] - input_data_2['Sirkette_Calisilan_Sure'])

# Yeni değişken türetme -9

input_data_2['Ortalama_Memnuniyet'] = (input_data_2['Ise_Dahiliyet'] + input_data_2['Is_Memnuniyeti'] + input_data_2['Ortam_Memnuniyeti'] + input_data_2['Iliski_Memnuniyeti'] + input_data_2['Is_Hayat_Dengesi'])/5

# Yeni değişken türetme -10

def Genc_Olmayan_Bekar(data) :
    if data['Yas'] > 35 and data['Medeni_Durum'] == 'Single':
        return 1
    else:
        return 0
input_data_2['Genc_Olmayan_Bekar'] = input_data_2.apply(lambda input_data_2:Genc_Olmayan_Bekar(input_data_2) ,axis = 1)

# Yeni değişken türetme -11
```

```
def Cok_Genc_Evli(data) :
    if data['Yas'] < 28 and data['Medeni_Durum'] == 'Married':
        return 1
    else:
        return 0
input_data_2['Cok_Genc_Evli'] = input_data_2.apply(lambda input_data_2:Cok_Genc_Evli(input_data_2) ,axis = 1)

# Yeni değişken türetme -12

def Az_Kazanan_Evli(data) :
    if data['Aylik_Gelir'] < 3500 and data['Medeni_Durum'] == 'Married':
        return 1
    else:
        return 0
input_data_2['Az_Kazanan_Evli'] = input_data_2.apply(lambda input_data_2:Az_Kazanan_Evli(input_data_2) ,axis = 1)

# Yeni değişken türetme -13

def Az_Kazanan_Bekar(data) :
    if data['Aylik_Gelir'] <3500 and (data['Medeni_Durum'] == 'Single' or data['Medeni_Durum'] == 'Divorced'):
        return 1
    else:
```

```
        return 0
    input_data_2['Az_Kazanan_Bekar'] = input_data_2.apply(lambda input_data_2:Az_Kazanan_Bekar(input_data_2) ,axis = 1)

    # Yeni değişken türetme -14

    def Cok_Kazanan_Bekar(data) :
      if data['Aylik_Gelir'] > 6500 and (data['Medeni_Durum'] == 'Single' or data['Medeni_Durum'] == 'Divorced'):
        return 1
      else:
        return 0
    input_data_2['Cok_Kazanan_Bekar'] = input_data_2.apply(lambda input_data_2:Cok_Kazanan_Bekar(input_data_2) ,axis = 1)

    # Yeni değişken türetme -15

    def Sene_Basina_Egitim_Sayisi(data) :
      if data['Sirkette_Calisilan_Sure'] >0:
        return (data['Egitim_Sayisi']/data['Sirkette_Calisilan_Sure'])
      else:
        return 0
    input_data_2['Sene_Basina_Egitim_Sayisi'] = input_data_2.apply(lambda input_data_2:Sene_Basina_Egitim_Sayisi(input_data_2) ,axis = 1)
```

```
input_data_derived_cols=['Terfi_Etmis', 'Rol_Degistirmis', 'Yonetici_Degistirmis',
'Is_Degistirmis', 'Is_Degistirme_Suresi',
                         'Is_Degistirme_Suresine_Yakinlik', 'Ise_Baslama_Yasi',
'Sirkete_Baslama_Yasi', 'Ortalama_Memnuniyet',
                         'Genc_Olmayan_Bekar', 'Cok_Genc_Evli', 'Az_Kazanan_Evli',
'Az_Kazanan_Bekar', 'Cok_Kazanan_Bekar',
                         'Sene_Basina_Egitim_Sayisi', 'Isten_Ayrilma']

input_data_derived_columns=input_data_2[input_data_derived_cols]


# Yeni değişken türetme - Yeni değişkenlerin analizi

num_col_to_draw_deriv = []
for column in input_data_derived_columns.columns:
    if input_data_derived_columns[column].dtypes != object:
        num_col_to_draw_deriv.append(column)

num_col_to_draw_deriv.remove("Isten_Ayrilma")

plt.figure(figsize=(20, 20))

for i, column in enumerate(num_col_to_draw_deriv, 1):
    plt.subplot(6, 4, i)
    input_data_derived_columns[input_data_derived_columns["Isten_Ayrilma"] ==
0][column].hist(bins=35, color='blue', label='Isten_Ayrilma = HAYIR', alpha=0.9)
```

```
        input_data_derived_columns[input_data_derived_columns["Isten_Ayrilma"] == 1][column].hist(bins=35, color='deeppink', label='Isten_Ayrilma = EVET', alpha=0.9)
        plt.legend()
        plt.xlabel(column)


    #Hedef değişken ile girdi değişkenlerin korelasyon analizi
    input_data_derived_columns.drop('Isten_Ayrilma', axis=1).corrwith(input_data_derived_columns.Isten_Ayrilma).plot(kind='barh', figsize=(10, 7))


    #İşten Ayrılma Durumuna Göre Sayısal Değişkenler için Dağılım Kontrolü

    for i, column in enumerate(num_col_to_draw_deriv, 1):
        (mu, sigma) = norm.fit(input_data_derived_columns.loc[input_data['Isten_Ayrilma'] == 0, column])
        print(column + ' İşten Ayrılma = HAYIR: Ortalama = {:.1f} ve Standart Sapma = {:.1f}'.format(mu, sigma))
        (mu, sigma) = norm.fit(input_data_derived_columns.loc[input_data['Isten_Ayrilma'] == 1, column])
        print(column + ' İşten Ayrılma = EVET: Ortalama = {:.1f} ve Standart Sapma = {:.1f}'.format(mu, sigma))


    #İşten ayrılma durumu bazında değişkenlerin dağılımı
```

```
plt.figure(figsize=(20, 40))

for i, column in enumerate(num_col_to_draw_deriv, 1):
    plt.subplot(12, 2, i)
    plt.style.use('seaborn-white')
    plt.grid(True, linestyle="-", color='grey', alpha=0.7)

sns.kdeplot(input_data_derived_columns.loc[input_data_derived_columns['Isten_Ayrilma'] == 0, column], color='blue', label='İşten Ayrılma = HAYIR', alpha=0.7)

sns.kdeplot(input_data_derived_columns.loc[input_data_derived_columns['Isten_Ayrilma'] == 1, column], color='deeppink', label='İşten Ayrılma = EVET', alpha=0.7)
    plt.rcParams['lines.linewidth'] = 2
    plt.xlabel(column)
    plt.ylabel('Çalışan Sayısı')
    plt.title(column + ' İşten ayrılma durumu bazında değişkenlerin dağılımı')
    plt.legend(loc='upper right', fancybox=True, framealpha=1, shadow=False, borderpad=1)
    plt.tight_layout();


object_col_to_draw

input_data_2[object_col_to_draw]

#Encoding
```

```
from sklearn.preprocessing import LabelEncoder

input_data_3=input_data_2.copy()

label = LabelEncoder()
for column in object_col_to_draw:
    input_data_3[column] = label.fit_transform(input_data_3[column])

#Train-test split


from sklearn.model_selection import train_test_split

X = input_data_3.drop('Isten_Ayrilma', axis=1)
y = input_data_3.Isten_Ayrilma

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, roc_auc_score, confusion_matrix, classification_report

def print_score(clf, X_train, y_train, X_test, y_test, train=True):
    if train:
        y_train_pred = clf.predict(X_train)
        clf_report = pd.DataFrame(classification_report(y_train, y_train_pred, output_dict=True))
```

```
            print("Train
Result:\n================================================")
            print(f"Accuracy Score: {accuracy_score(y_train, y_train_pred) * 100:.2f}%")
            print("_______________________________________________")
            print(f"Precision Score: {precision_score(y_train, y_train_pred, average='weighted') *
100:.2f}%")
            print("_______________________________________________")
            print(f"Recall Score: {recall_score(y_train, y_train_pred, average='weighted') *
100:.2f}%")
            print("_______________________________________________")
            print(f"F1 Score: {f1_score(y_train, y_train_pred, average='weighted') * 100:.2f}%")
            print("_______________________________________________")
            print(f"CLASSIFICATION REPORT:\n{clf_report}")
            print("_______________________________________________")
            print(f"Confusion Matrix: \n {confusion_matrix(y_train, y_train_pred)}\n")

        elif train==False:
            y_test_pred = clf.predict(X_test)
            clf_report = pd.DataFrame(classification_report(y_test, y_test_pred, output_dict=True))
            print("Test
Result:\n================================================")
            print(f"Accuracy Score: {accuracy_score(y_test, y_test_pred) * 100:.2f}%")
            print("_______________________________________________")
            print(f"Precision Score: {precision_score(y_test, y_test_pred, average='weighted') *
100:.2f}%")
            print("_______________________________________________")
```

```
        print(f"Recall Score: {recall_score(y_test, y_test_pred, average='weighted') * 100:.2f}%")
        print("_____")
        print(f"F1 Score: {f1_score(y_test, y_test_pred, average='weighted') * 100:.2f}%")
        print("_____")
        print(f"CLASSIFICATION REPORT:\n{clf_report}")
        print("_____")
        print(f"Confusion Matrix: \n {confusion_matrix(y_test, y_test_pred)}\n")


#Decision Tree Classifier


from sklearn.tree import DecisionTreeClassifier


dtree_clf = DecisionTreeClassifier(random_state=42)
dtree_clf.fit(X_train, y_train)


print_score(dtree_clf, X_train, y_train, X_test, y_test, train=True)
print_score(dtree_clf, X_train, y_train, X_test, y_test, train=False)


importances = dtree_clf.feature_importances_
feat_importances = pd.Series(importances, index=X_test.columns)
indices = feat_importances.argsort(feat_importances)
indice_names = [list(X_test.columns)[i] for i in indices]
plt.figure(figsize=(10,10))
```

```
plt.barh(indice_names, feat_importances[indices])
plt.xlabel('Önem derecesi')
plt.ylabel('Değişken')
plt.title('Değişken önem dereceleri')
plt.show()

#confusion matrix
y_test_pred = dtree_clf.predict(X_test)
fig = plt.figure(figsize=(5,5))
cm = confusion_matrix(y_test, y_test_pred)
ax = plt.subplot()
sns.heatmap(cm, annot=True, ax=ax, fmt='g', cmap='PuBu')
plt.xlabel('Tahmin')
plt.ylabel('Gerçekleşen')
plt.title('Hata matrisi')
plt.show()



#Random Forest

from sklearn.ensemble import RandomForestClassifier

rf_clf = RandomForestClassifier(n_estimators=100)
rf_clf.fit(X_train, y_train)
```

```
print_score(rf_clf, X_train, y_train, X_test, y_test, train=True)

print_score(rf_clf, X_train, y_train, X_test, y_test, train=False)


importances = rf_clf.feature_importances_

feat_importances = pd.Series(importances, index=X_test.columns)

indices = feat_importances.argsort(feat_importances)

indice_names = [list(X_test.columns)[i] for i in indices]

plt.figure(figsize=(10,10))

plt.barh(indice_names, feat_importances[indices])

plt.xlabel('Önem derecesi')

plt.ylabel('Değişken')

plt.title('Değişken önem dereceleri')

plt.show()


#confusion matrix

y_test_pred = rf_clf.predict(X_test)

fig = plt.figure(figsize=(5,5))

cm = confusion_matrix(y_test, y_test_pred)

ax = plt.subplot()

sns.heatmap(cm, annot=True, ax=ax, fmt='g', cmap='PuBu')

plt.xlabel('Tahmin')

plt.ylabel('Gerçekleşen')

plt.title('Hata matrisi')

plt.show()
```

```
#Logistic Regression

from sklearn.linear_model import LogisticRegression

lr_clf = LogisticRegression(solver='liblinear', penalty='l1')
lr_clf.fit(X_train, y_train)

print_score(lr_clf, X_train, y_train, X_test, y_test, train=True)
print_score(lr_clf, X_train, y_train, X_test, y_test, train=False)

importances = abs(np.std(X_train,0)*lr_clf.coef_[0])
feat_importances = pd.Series(importances, index=X_test.columns)
indices = feat_importances.argsort(feat_importances)
indice_names = [list(X_test.columns)[i] for i in indices]
plt.figure(figsize=(10,10))
plt.barh(indice_names, feat_importances[indices])
plt.xlabel('Önem derecesi')
plt.ylabel('Değişken')
plt.title('Değişken önem dereceleri')
plt.show()

#confusion matrix
y_test_pred = lr_clf.predict(X_test)
fig = plt.figure(figsize=(5,5))
```

```
cm = confusion_matrix(y_test, y_test_pred)
ax = plt.subplot()
sns.heatmap(cm, annot=True, ax=ax, fmt='g', cmap='PuBu')
plt.xlabel('Tahmin')
plt.ylabel('Gerçekleşen')
plt.title('Hata matrisi')
plt.show()

#Support Vector Machine

from sklearn.svm import SVC

svm_clf = SVC(kernel='linear')
svm_clf.fit(X_train, y_train)

print_score(svm_clf, X_train, y_train, X_test, y_test, train=True)
print_score(svm_clf, X_train, y_train, X_test, y_test, train=False)


importances = svm_clf.feature_importances_
feat_importances = pd.Series(importances, index=X_test.columns)
indices = feat_importances.argsort(feat_importances)
indice_names = [list(X_test.columns)[i] for i in indices]
plt.figure(figsize=(10,10))
plt.barh(indice_names, feat_importances[indices])
```

```
plt.xlabel('Önem derecesi')
plt.ylabel('Değişken')
plt.title('Değişken önem dereceleri')
plt.show()


#confusion matrix
y_test_pred = svm_clf.predict(X_test)
fig = plt.figure(figsize=(5,5))
cm = confusion_matrix(y_test, y_test_pred)
ax = plt.subplot()
sns.heatmap(cm, annot=True, ax=ax, fmt='g', cmap='PuBu')
plt.xlabel('Tahmin')
plt.ylabel('Gerçekleşen')
plt.title('Hata matrisi')
plt.show()


#XGBoost

from xgboost import XGBClassifier

xgb_clf = XGBClassifier()
xgb_clf.fit(X_train, y_train)
```

```
print_score(xgb_clf, X_train, y_train, X_test, y_test, train=True)

print_score(xgb_clf, X_train, y_train, X_test, y_test, train=False)


importances = xgb_clf.feature_importances_

feat_importances = pd.Series(importances, index=X_test.columns)

indices = feat_importances.argsort(feat_importances)

indice_names = [list(X_test.columns)[i] for i in indices]

plt.figure(figsize=(10,10))

plt.barh(indice_names, feat_importances[indices])

plt.xlabel('Önem derecesi')

plt.ylabel('Değişken')

plt.title('Değişken önem dereceleri')

plt.show()


#confusion matrix

y_test_pred = xgb_clf.predict(X_test)

fig = plt.figure(figsize=(5,5))

cm = confusion_matrix(y_test, y_test_pred)

ax = plt.subplot()

sns.heatmap(cm, annot=True, ax=ax, fmt='g', cmap='PuBu')

plt.xlabel('Tahmin')

plt.ylabel('Gerçekleşen')

plt.title('Hata matrisi')

plt.show()
```

```
#AdaBoost

from sklearn.ensemble import AdaBoostClassifier

ab_clf = AdaBoostClassifier()
ab_clf.fit(X_train, y_train)

print_score(ab_clf, X_train, y_train, X_test, y_test, train=True)
print_score(ab_clf, X_train, y_train, X_test, y_test, train=False)

importances = ab_clf.feature_importances_
feat_importances = pd.Series(importances, index=X_test.columns)
indices = feat_importances.argsort(feat_importances)
indice_names = [list(X_test.columns)[i] for i in indices]
plt.figure(figsize=(10,10))
plt.barh(indice_names, feat_importances[indices])
plt.xlabel('Önem derecesi')
plt.ylabel('Değişken')
plt.title('Değişken önem dereceleri')
plt.show()

#confusion matrix
y_test_pred = ab_clf.predict(X_test)
fig = plt.figure(figsize=(5,5))
cm = confusion_matrix(y_test, y_test_pred)
```

```
ax = plt.subplot()
sns.heatmap(cm, annot=True, ax=ax, fmt='g', cmap='PuBu')
plt.xlabel('Tahmin')
plt.ylabel('Gerçekleşen')
plt.title('Hata matrisi')
plt.show()

#CatBoost

from catboost import CatBoostClassifier

cb_clf = CatBoostClassifier()
cb_clf.fit(X_train, y_train, verbose=0)

print_score(cb_clf, X_train, y_train, X_test, y_test, train=True)
print_score(cb_clf, X_train, y_train, X_test, y_test, train=False)

importances = cb_clf.feature_importances_
feat_importances = pd.Series(importances, index=X_test.columns)
indices = feat_importances.argsort(feat_importances)
indice_names = [list(X_test.columns)[i] for i in indices]
plt.figure(figsize=(10,10))
plt.barh(indice_names, feat_importances[indices])
plt.xlabel('Önem derecesi')
plt.ylabel('Değişken')
```

```
plt.title('Değişken önem dereceleri')
plt.show()

#confusion matrix
y_test_pred = cb_clf.predict(X_test)
fig = plt.figure(figsize=(5,5))
cm = confusion_matrix(y_test, y_test_pred)
ax = plt.subplot()
sns.heatmap(cm, annot=True, ax=ax, fmt='g', cmap='PuBu')
plt.xlabel('Tahmin')
plt.ylabel('Gerçekleşen')
plt.title('Hata matrisi')
plt.show()


#LightGBM

from lightgbm import LGBMClassifier

lgb_clf = LGBMClassifier()
lgb_clf.fit(X_train, y_train)

print_score(lgb_clf, X_train, y_train, X_test, y_test, train=True)
print_score(lgb_clf, X_train, y_train, X_test, y_test, train=False)
```

```
importances = lgb_clf.feature_importances_
feat_importances = pd.Series(importances, index=X_test.columns)
indices = feat_importances.argsort(feat_importances)
indice_names = [list(X_test.columns)[i] for i in indices]
plt.figure(figsize=(10,10))
plt.barh(indice_names, feat_importances[indices])
plt.xlabel('Önem derecesi')
plt.ylabel('Değişken')
plt.title('Değişken önem dereceleri')
plt.show()

#confusion matrix
y_test_pred = lgb_clf.predict(X_test)
fig = plt.figure(figsize=(5,5))
cm = confusion_matrix(y_test, y_test_pred)
ax = plt.subplot()
sns.heatmap(cm, annot=True, ax=ax, fmt='g', cmap='PuBu')
plt.xlabel('Tahmin')
plt.ylabel('Gerçekleşen')
plt.title('Hata matrisi')
plt.show()
```

```
ml_models = {
    'Decision Tree': dtree_clf,
    'Random Forest': rf_clf,
    'XGBoost': xgb_clf,
    'Logistic Regression': lr_clf,
    'Support Vector Machine': svm_clf,
    'LightGBM': lgb_clf,
    'CatBoost': cb_clf,
    'AdaBoost': ab_clf
}

for model in ml_models:
    print(f"{model.upper():{30}}  roc_auc_score:  {roc_auc_score(y_test, ml_models[model].predict(X_test)):.3f}")


importances = abs(np.std(X_train,0)*lr_clf.coef_[0])
feat_importances = pd.Series(importances, index=X_test.columns)
indices = feat_importances.argsort(feat_importances)
indice_names = [list(X_test.columns)[i] for i in indices]
plt.figure(figsize=(10,10))
plt.barh(indice_names, feat_importances[indices])
plt.xlabel('Önem derecesi')
plt.ylabel('Değişken')
plt.title('Değişken önem dereceleri')
```

```
plt.show()



#Logistic regression features and coeefficients

log_reg_df1 = abs(np.std(X_train,0))/abs(np.std(X_train,0))*lr_clf.coef_[0]

log_reg_df1.to_excel("C:/Users/Asus/Desktop/KadirYildiz/Thesis/Output/Lojistik_Regresyon_Formulu.xlsx")


Sensitivity_data=X_train.append(X_test)


#Model objesini dışarı çıkarma


import pickle

pickle.dump(dtree_clf, open("C:/Users/Asus/Desktop/KadirYildiz/Thesis/Output/model_dtree.pkl", 'wb'))

pickle.dump(rf_clf, open("C:/Users/Asus/Desktop/KadirYildiz/Thesis/Output/model_rf.pkl", 'wb'))

pickle.dump(xgb_clf, open("C:/Users/Asus/Desktop/KadirYildiz/Thesis/Output/model_xgb.pkl", 'wb'))

pickle.dump(lr_clf, open("C:/Users/Asus/Desktop/KadirYildiz/Thesis/Output/model_lr.pkl", 'wb'))

pickle.dump(svm_clf, open("C:/Users/Asus/Desktop/KadirYildiz/Thesis/Output/model_svm.pkl", 'wb'))

pickle.dump(lgb_clf, open("C:/Users/Asus/Desktop/KadirYildiz/Thesis/Output/model_lgb.pkl", 'wb'))

pickle.dump(cb_clf, open("C:/Users/Asus/Desktop/KadirYildiz/Thesis/Output/model_cb.pkl", 'wb'))
```

```
pickle.dump(ab_clf,
open("C:/Users/Asus/Desktop/KadirYildiz/Thesis/Output/model_ab.pkl", 'wb'))

import pandas as pd

import numpy as np


path = 'C:/Users/Asus/Desktop/KadirYildiz/Thesis'

path_input = 'C:/Users/Asus/Desktop/KadirYildiz/Thesis/Data'

path_output = 'C:/Users/Asus/Desktop/KadirYildiz/Thesis/Output'


#read the input dataset

Sensitivity_data = pd.read_excel(path_output + "/" + "Sensitivity_data.xlsx")


Sensitivity_data.head()


Sensitivity_data.rename( columns={'Unnamed: 0':'ID'}, inplace=True )


Sensitivity_data = Sensitivity_data.sort_values('ID')


#load the light gbm model from disk

import pickle

loaded_model                                                                    =
pickle.load(open("C:/Users/Asus/Desktop/KadirYildiz/Thesis/Output/model_lgb.pkl", 'rb'))


Sensitivity_data = Sensitivity_data.drop('ID', axis=1)


Sensitivity_data_case00 = Sensitivity_data.copy()
```

```
Sensitivity_data_case00_preds = loaded_model.predict_proba(Sensitivity_data_case00)

Sensitivity_data_case00_preds=Sensitivity_data_case00_preds[:,1]

Sensitivity_data_case00['preds'] = Sensitivity_data_case00_preds

###%1 maaş artışı

Sensitivity_data_case01 = Sensitivity_data.copy()

Sensitivity_data_case01['Aylik_Gelir'] =
round(Sensitivity_data_case01['Aylik_Gelir']*(1.01))

Sensitivity_data_case01_preds = loaded_model.predict_proba(Sensitivity_data_case01)

Sensitivity_data_case01_preds=Sensitivity_data_case01_preds[:,1]

Sensitivity_data_case01['preds'] = Sensitivity_data_case01_preds

print("Impact of 1% salary increase on turnover is: " +
str(round((Sensitivity_data_case01['preds'].mean()/Sensitivity_data_case00['preds'].mean()-
1)*100,2)) + "%")

print("Impact of 1% salary increase is: " +
```

```
        str(Sensitivity_data_case01['Aylik_Gelir'].sum() -
Sensitivity_data_case00['Aylik_Gelir'].sum()) + "USD")

        ###%3 maaş artışı

        Sensitivity_data_case03 = Sensitivity_data.copy()

        Sensitivity_data_case03['Aylik_Gelir'] =
round(Sensitivity_data_case03['Aylik_Gelir']*(1.03))

        Sensitivity_data_case03_preds = loaded_model.predict_proba(Sensitivity_data_case03)

        Sensitivity_data_case03_preds=Sensitivity_data_case03_preds[:,1]

        Sensitivity_data_case03['preds'] = Sensitivity_data_case03_preds

        print("Impact of 3% salary increase on turnover is: " +
str(round((Sensitivity_data_case03['preds'].mean()/Sensitivity_data_case00['preds'].mean()-
1)*100,2)) + "%")

        print("Impact of 3% salary increase is: " +
          str(Sensitivity_data_case03['Aylik_Gelir'].sum() -
Sensitivity_data_case00['Aylik_Gelir'].sum()) + "USD")

        ###%5 maaş artışı
```

```
Sensitivity_data_case05 = Sensitivity_data.copy()

Sensitivity_data_case05['Aylik_Gelir'] = round(Sensitivity_data_case05['Aylik_Gelir']*(1.05))

Sensitivity_data_case05_preds = loaded_model.predict_proba(Sensitivity_data_case05)

Sensitivity_data_case05_preds=Sensitivity_data_case05_preds[:,1]

Sensitivity_data_case05['preds'] = Sensitivity_data_case05_preds

print("Impact of 5% salary increase on turnover is: " +

str(round((Sensitivity_data_case05['preds'].mean()/Sensitivity_data_case00['preds'].mean()-1)*100,2)) + "%")

print("Impact of 5% salary increase is: " +

   str(Sensitivity_data_case05['Aylik_Gelir'].sum() - Sensitivity_data_case00['Aylik_Gelir'].sum()) + "USD")

###%7 maaş artışı

Sensitivity_data_case07 = Sensitivity_data.copy()
```

```
Sensitivity_data_case07['Aylik_Gelir'] = round(Sensitivity_data_case07['Aylik_Gelir']*(1.07))

Sensitivity_data_case07_preds = loaded_model.predict_proba(Sensitivity_data_case07)

Sensitivity_data_case07_preds=Sensitivity_data_case07_preds[:,1]

Sensitivity_data_case07['preds'] = Sensitivity_data_case07_preds

print("Impact of 7% salary increase on turnover is: " +
str(round((Sensitivity_data_case07['preds'].mean()/Sensitivity_data_case00['preds'].mean()-1)*100,2)) + "%")

print("Impact of 7% salary increase is: " +
   str(Sensitivity_data_case07['Aylik_Gelir'].sum() - Sensitivity_data_case00['Aylik_Gelir'].sum()) + "USD")

###%9 maaş artışı

Sensitivity_data_case09 = Sensitivity_data.copy()

Sensitivity_data_case09['Aylik_Gelir'] = round(Sensitivity_data_case09['Aylik_Gelir']*(1.09))

Sensitivity_data_case09_preds = loaded_model.predict_proba(Sensitivity_data_case09)
```

```
Sensitivity_data_case09_preds=Sensitivity_data_case09_preds[:,1]

Sensitivity_data_case09['preds'] = Sensitivity_data_case09_preds

print("Impact of 9% salary increase on turnover is: " +

str(round((Sensitivity_data_case09['preds'].mean()/Sensitivity_data_case00['preds'].mean()-
1)*100,2)) + "%")

print("Impact of 9% salary increase is: " +
   str(Sensitivity_data_case09['Aylik_Gelir'].sum() -
Sensitivity_data_case00['Aylik_Gelir'].sum()) + "USD")

###%11 maaş artışı

Sensitivity_data_case11 = Sensitivity_data.copy()

Sensitivity_data_case11['Aylik_Gelir'] =
round(Sensitivity_data_case11['Aylik_Gelir']*(1.11))

Sensitivity_data_case11_preds = loaded_model.predict_proba(Sensitivity_data_case11)

Sensitivity_data_case11_preds=Sensitivity_data_case11_preds[:,1]

Sensitivity_data_case11['preds'] = Sensitivity_data_case11_preds
```

```
print("Impact of 11% salary increase on turnover is: " +

str(round((Sensitivity_data_case11['preds'].mean()/Sensitivity_data_case00['preds'].mean()-
1)*100,2)) + "%")

print("Impact of 11% salary increase is: " +
   str(Sensitivity_data_case11['Aylik_Gelir'].sum() -
Sensitivity_data_case00['Aylik_Gelir'].sum()) + "USD")

###%13 maaş artışı

Sensitivity_data_case13 = Sensitivity_data.copy()

Sensitivity_data_case13['Aylik_Gelir'] =
round(Sensitivity_data_case13['Aylik_Gelir']*(1.13))

Sensitivity_data_case13_preds = loaded_model.predict_proba(Sensitivity_data_case13)

Sensitivity_data_case13_preds=Sensitivity_data_case13_preds[:,1]

Sensitivity_data_case13['preds'] = Sensitivity_data_case13_preds

print("Impact of 13% salary increase on turnover is: " +
```

```
str(round((Sensitivity_data_case13['preds'].mean()/Sensitivity_data_case00['preds'].mean()-
1)*100,2)) + "%")

        print("Impact of 13% salary increase is: " +
            str(Sensitivity_data_case13['Aylik_Gelir'].sum() -
Sensitivity_data_case00['Aylik_Gelir'].sum()) + "USD")

        ###%15 maaş artışı

        Sensitivity_data_case15 = Sensitivity_data.copy()

        Sensitivity_data_case15['Aylik_Gelir'] =
round(Sensitivity_data_case15['Aylik_Gelir']*(1.15))

        Sensitivity_data_case15_preds = loaded_model.predict_proba(Sensitivity_data_case15)

        Sensitivity_data_case15_preds=Sensitivity_data_case15_preds[:,1]

        Sensitivity_data_case15['preds'] = Sensitivity_data_case15_preds

        print("Impact of 15% salary increase on turnover is: " +

str(round((Sensitivity_data_case15['preds'].mean()/Sensitivity_data_case00['preds'].mean()-
1)*100,2)) + "%")
```

```
print("Impact of 15% salary increase is: " +
    str(Sensitivity_data_case15['Aylik_Gelir'].sum() -
Sensitivity_data_case00['Aylik_Gelir'].sum()) + "USD")


#read the input dataset
optimization_data = pd.read_excel(path_output + "/" +
"merged_data_for_optimization.xlsx")

optimization_data['ID'] = optimization_data['Unnamed: 0']+1

optimization_data['ID'] = "Employee_Nr_" + optimization_data['ID'].map(str)

M = ['preds_00', 'preds_01', 'preds_03', 'preds_05', 'preds_07', 'preds_09', 'preds_11',
'preds_13', 'preds_15']

S = list(optimization_data['Aylik_Gelir'])

P = list(optimization_data['Performans_Notu'])

T_yeni= optimization_data['Aylik_Gelir'].sum()*1.10

def wage_optimizer_model(N, M, P, A, S):
    model = pyo.ConcreteModel(name="(wage_optimizer_model)")

    model.d = pyo.Var(N, M, within=pyo.Binary)
```

```
    def obj_rule(mdl):
        return sum(P[n,m]*(1-A[n,m])*mdl.d[n,m] for n in N for m in M)
    model.obj = pyo.Objective(rule=obj_rule, sense=pyo.maximize)

    def tek_maas_kisiti(mdl, n):
        return sum(mdl.d[n,m] for m in M) == 1
    model.tek_maas_kisiti = pyo.Constraint(N, rule=tek_maas_kisiti)

    def butce_kisiti(mdl):
        return sum(S[n,m]*mdl.d[n,m] for n in N for m in M) <= T_yeni
    model.butce_kisiti = pyo.Constraint(rule=butce_kisiti)

    return model

# Create the Pyomo model
model = wage_optimizer_model(N, M, P, A, S)

# Create the solver interface and solve the model
solver = pyo.SolverFactory('glpk')
res = solver.solve(model)

pyo.assert_optimal_termination(res)

model.d.pprint()
```

```
# solver results

res.write()
```