HACETTEPE ÜNİVERSİTESİ

İST434 Büyük Veri Analitiği Dersi Final Ödevi
Doç. Dr. Duygu İçen

Konu: Makine Öğrenimi Yöntemleri ile Çalışan Kayıp Tahmini

21821665 - Ecem ÇIRAKOĞLU

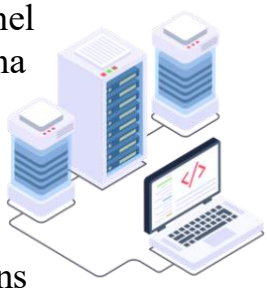
Haziran 2023

İçindekiler Tablosu

Büyük Veri ve İstatistik

Büyük veri ve istatistik birbirlerine göre farklılıkları da olan iki ilişkili kavramdır. **İstatistik**, kitleden kurallara uygun olarak yeterli sayıda verilerin toplanması, düzenlenmesi, çözümlenmesi ve sonuçlarının yorumlanması ile ilgili yöntemleri içeren bir bilim dalı iken, **büyük veri**, boyutu hızla büyüyen, farklı formatta, depolanması, işlenmesi ve analiz edilmesi geleneksel yöntemler ile mümkün olmayan veridir.

Her iki kavram da sağlık, bankacılık, sigortacılık, e-ticaret vb. gibi çeşitli alanlar için verileri analiz etmekte, amaca yönelik ortaya anlamlı sonuçlar çıkartmakta ve böylece bu alanlara yönelik karar vericiye karar verme sürecinde katkı sağlamaktadırlar.



Ancak bu iki kavram farklılıklar da göstermektedir. En temel farklılıklardan birisi kullanılan veri hacmidir. İstatistik daha küçük düzenli yapılandırılmış veriler üzerinde çalışırken, büyük veri çeşitli kaynaklardan toplanan büyük hacimli yapılandırılmış, yapılandırılmamış veya yarı-yapılandırılmış veriler üzerinde çalışmaktadır.
Bir diğer farklılık ise istatistik, hipotez testleri, tek yönlü varyans analizi, gibi istatistiksel yöntemler kullanmakta iken, büyük verilerde tüm veri ile uğraşıldığından hipotez testi durumu söz konusu değildir. Genellikle ölçeklenebilir, hata toleransı yüksek dağıtık platformlar ve makine öğrenimi algoritmaları kullanılmaktadır.

Bunu bir örnek ile açıklamak gerekirse, bir e-ticaret sitesinden alışveriş yapan müşterinin davranışlarını tahmin etmek adına bir çalışma yaptığımızı düşünelim. Bu durumda istatistik, müşterinin işlemlerinden oluşan yapılandırılmış küçük bir veri kümesini analiz ederek müşteri popülasyonu hakkında istatistiksel yöntemler ile anlamlı çıkarımlar elde etmek anlamına gelmektedir. Büyük veri ise, müşterilerin davranışlarının daha eksiksik bir şekilde ortaya çıkartmak için tüm müşteri alışveriş işlemlerini, ürün yorumlarını ve diğer bütün veri kaynaklarını içerebilecek büyük veri kümesini Hadoop, Spark gibi dağıtık platformlar ve makine öğrenimi algoritmaları ile analiz ederek müşteri popülasyonu hakkında anlamlı çıkarımlar elde etmektedir.

Problem Tanımı

Çalışan kaybı, şirketler için önem arz eden bir durumdur. Çalışan kaybı maliyetlerin artmasına, iş yükünün artmasına ve yetenekli insanların kaybedilmesine yol açabilmektedir. Böylece şirketin performansını kötü yönde etkilenebilmektedir. Bu durumların önüne geçebilmek için şirketler, çalışan kaybına neden olan faktörleri belirlemek ve buna bağlı olarak çalışanları elde tutmaya yönelik stratejiler geliştirmelidir.
Bu çalışmanın amacı, çalışan kaybına neden olan faktörleri belirlemek ve işten çıkma olasılığı yüksek olan çalışanları doğru bir şekilde tahmin etmektir. Böylelikle şirketler işten ayrılma olasılığı yüksek çalışanlara yönelik müdehalelerde bulunabilir.

Çalışma, şirketlerin satış, araştırma ve geliştirme, insan kaynakları departmanında bulunan kişiler üzerine yapılmıştır. Hedef değişken, 0 (çalışanın kalması) ve 1 (çalışanın ayrılması) olmak üzere ikili bir değişken olan 'Attrition' değişkenidir.

https://whatfix.com/blog/employee-churn/

Veri Seti Hakkında

Bu çalışmada Kaggle'da bulunan ve IBM veri bilimcileri tarafından oluşturulan yapay veri seti kullanılmıştır. Veri seti çalışanların demografik bilgileri, eğitim durumları, maaşları, iş rolleri vb. gibi bilgileri çermektedir. Veri setinde toplam 1470 çalışan ve çalışanlara ait 35 değişken bulunmaktadır. Bu değişkenler Tablo 1'de listelenmiştir.

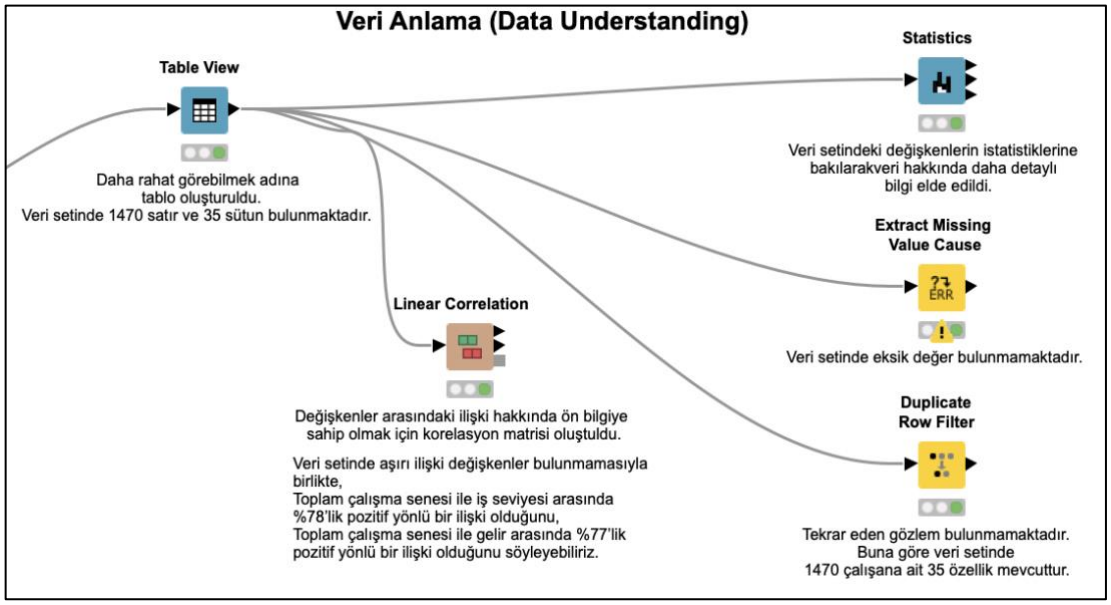
Tablo 1

Age: Yaş
Attrition (Hedef değişken): Çalışanın işte ayrılma-ayrılamam durumu (ayrılan çalışan 1, kalan çalışan 0)
BusinessTravel: Çalışanın seyat etme sıklığı
DailyRate: Çalışanın günlük ücreti
Department: Çalışanın bulunduğu departman (Satış, Araştırma ve Geliştirme, İnsan Kaynakları)
DistanceFromHome: Çalışanın evi ile işi arasındaki mesafe
Education: Çalışanın eğitim durumu (1:Ortaokul 2:Lise 3:Lisans 4:Yüksek Lisans 5:Doktor)
EducationField: Eğitim Alanı
EmployeeCount: Çalışan sayısı
EmployeeNumber: Çalışan numarası
EnvironmentSatisfaction: Çevre memnuniyeti (1: Düşük, 2 :Orta, 3 :Yüksek, 4 :Çok Yüksek)
Gender: Cinsiyet
HourlyRate: Çalışanın saatlik ücreti
JobInvolment: İşe katılım (1: Düşük, 2:Orta, 3:Yüksek, 4:Çok Yüksek)
JobLevel: İş seviyesi
JobRole: İş rolü
JobSatisfaction: İş tatmini (1: Düşük, 2: Orta, 3 :Yüksek, 4 :Çok Yüksek)
MaritalStatus: Çalışanın medeni durumu
MonthlyIncome: Çalışanın aylık geliri
MonthlyRate: Çalışanın aylık maaşı
NumCompaniesWorked: Çalışanın mevcut şirketten önce çalıştığı şirket sayısı
Over18: Çalışanın 18 yaş üstü olup olmadığı
OverTime: Çalışanın fazladan mesai yapmadurumu
PercentSalaryHike: Maaş artış yüzdesi
PerformanceRating: Performans derecesi (1: Düşük, 2 :İyi, 3 :Mükemmel, 4 :Üstün)
RelationshipSatisfaction: Çalışanın ilişki memnuniyeti (1: Düşük, 2 :Orta, 3 :Yüksek, 4 :Çok Yüksek)
StandardHours: Her çalışan için standart çalışma saati (80 Saat)
StockOptionLevel: Çalışanın stok seviyesi
TotalWorkingYears: Çalışanın toplam çalışma yılı (0 ile 40 yıl)
TrainingTimesLastYear: Çalışanın son bir yıldaki eğitim süresi
WorkLifeBalance: Çalışanın iş-hayat dengesi (1:Kötü, 2 :İyi, 3 :Daha İyi, 4 :En İyi)
YearsAtCompany: Çalışanın şirketteki toplam çalışma yılıdır (0 ile 40 yıl)
YearsInCurrentRole: Çalışanın şirketteki mevcut pozisyonunda çalıştığı yıl(0 ile 18 yıl)
YearsSinceLastPromotion: Çalışanın en son terfi aldığı zaman (0 ile 15 yıl)
YearsWithCurrManager: Çalışanın mevcut yöneticisi ile çalıştığı süre (0 ile 17 yıl)

Uygulama

Çalışan kayıp tahmini çalışması, makine öğrenimi ve veri madenciliği gibi işlemleri kod yazmadan gerçekleştirebileceğimiz KNIME programı üzerinden yapılmıştır. Analiz için elimizdeki veri seti ile Şekil 1'de görüldüğü gibi birden çok aşamayı kapsayan bir akış oluşturulmuştur. Bu akış veri anlama, veri hazırlığı, veri görselleştirme, modelleme ve değerlendirme aşamalarını içermektedir.

Veri Anlama (Data Understanding): Bu aşama veri seti yapısını, değişkenlerini, değişkenler arasındaki ilişkileri anlamak için kullanılmıştır. Aynı zamanda değişkenlerde analiz öncesinde iyileştirme yapılma ihtiyacı olup olmadığına bakılmıştır.



Şekil 1

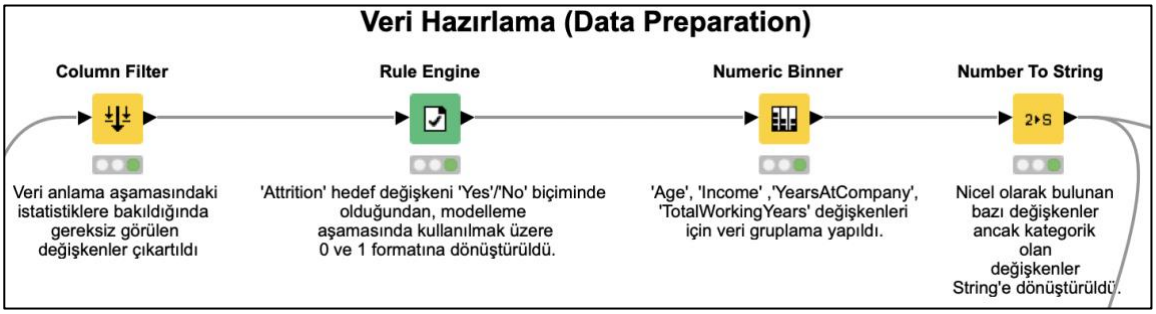
Bu aşamadan elde ettiğimiz sonuçlara göre,

- 1470 çalışan ve onlara ait 35 değişken bulunmaktadır,
- Eksik değer ve tekrar eden gözlemler bulunmamaktadır.
- Çalışan yaşının ortalama 36 olarak hesaplanmıştır.
- Çalışan maaşının ortalama 6.502 dolardır.
- Çalışanların toplam çalışma senesi ortalama 11 yıldır.

Aynı zamanda hangi değişkenler arasında ilişki olduğu hakkında ön bilgiye sahip olmak için oluşturulan korelasyon matrisine göre yüksek ve farklılık yaratabilecek ilişkilerin bulunmadığını ancak,

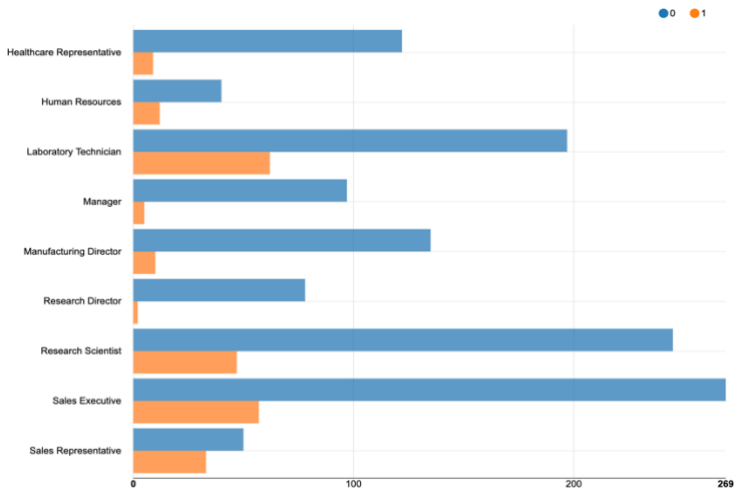
- Toplam çalışma senesi ile iş seviyesi arasında %78'lik; pozitif yönlü bir ilişki olduğunu,
- Toplam çalışma senesi ile gelir arasında %77'lik pozitif yönlü bir ilişki olduğunu söyleyebiliriz.

Veri Hazırlama (Data Preparation): Bir önceki aşamadan elde edinilen bilgilerden yola çıkarak veri seti üzerinde birtakım değişiklikler yapıldı. Örneğin, 'EmployeeCount', 'EmployeNumber' ve 'StandardHour' değişkenleri bütün gözlemlerde aynı değere sahip olduğu ve hedef değişken üzerinde etkisi olmayacağından çıkartıldı. Analizde kullanmak üzere hedef değişkeni 0 ve 1 formatına dönüştürüldü. Ardından analiz ve görselleştirme aşaması için faydası olabileceği düşünülerek yaş, gelir, şirkette çalışılan yıl sayısı ve toplam çalışılan yıl sayısı değişkenleri kendi içlerinde gruplandırıldı. Son olarak ise çalışanların eğitim durumu gibi kategorik olup veri setinde sayı olarak görülen değişkenler için format dönüşümü yapıldı.

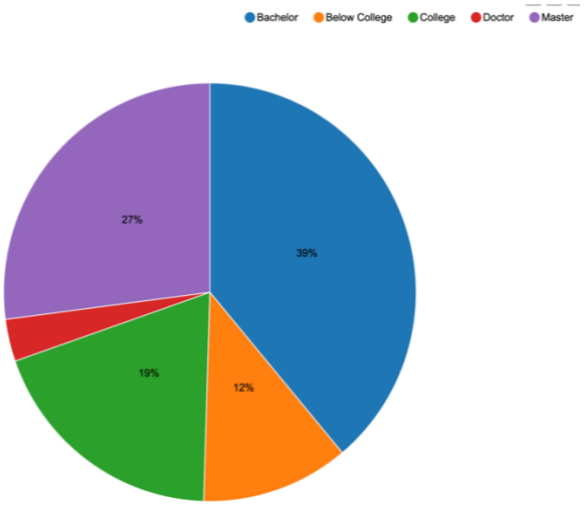


Şekil 2

Veri Görselleştirme (Data Visualization): Veriyi daha iyi anlamak, çalışanların eğilimlerini daha net görebilmek için görselleştirme aşaması oluşturulmuştur. Bar, histogram, pasta ve Sunburst grafikleri oluşturulmuştur.
(Aslında bu aşama veri hazırlama aşamasından önce yapılabilirdi ancak bazı nicel değişkenler, veri hazırlama aşamasında gruplandırılıp bu aşamada kullanılması düşünülmüştür.)



Şekil 3



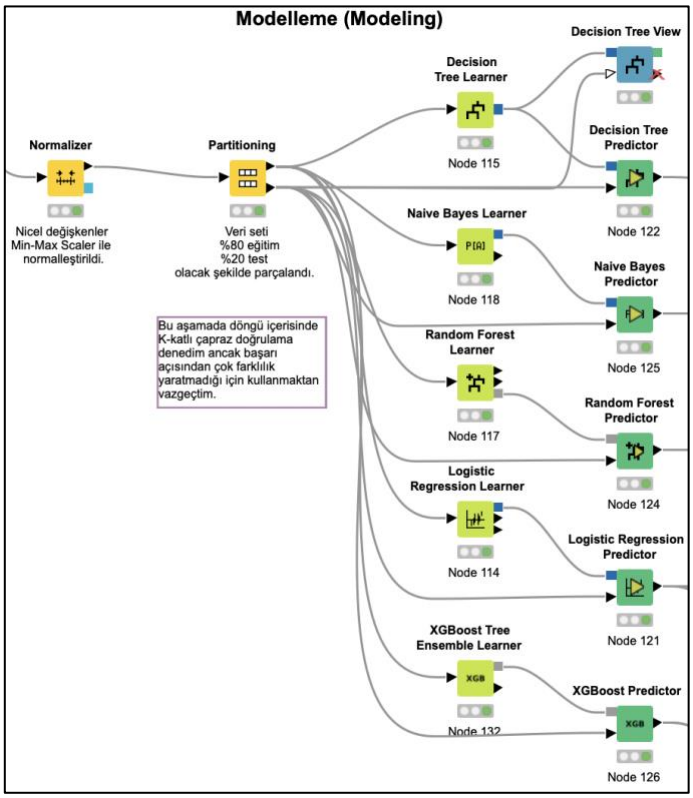
Şekil 4

Elde edilen grafiklere göre özetle,

- Veri setinde yer alan çalışanların %84'ünün işten ayrılmadığı; %16'sının işten ayrıldığı,
- Çalışanların eğitim durumunun %12'sinin ortaokul, %19'unun lise, %39'unun lisans, %27'sinin ise yüksek lisans olduğu,
- Araştırma ve geliştirme departmanında çalışanların diğer departmanlara göre işten ayrılma oranın daha fazla olduğu,
- 5 yıldan daha az süredir aynı yerde çalışan insanların işten çıkmalarının daha sık görüldüğü,

- Fazladan mesai kalmak zorunda olan çalışanların daha yüksek oranda işten ayrılmakta olduğu,
- Satış Yöneticisi ve laboratuvar teknisyeni olarak çalışanların, diğer iş rollerine göre daha önemli bir oranda işten ayrıldıkları söylenebilmektedir.
- Düşük maaş kategorisindeki (<6.500) maaşa sahip olan çalışanlardan işte kalanların oranının işten ayrılanlara göre daha fazla olduğunu; aynı şekilde daha yüksek maaş kategorisindeki maaşa sahip çalışanlardan da işte kalanların oranının işten ayrılanlara göre daha fazla olduğunu söyleyebiliriz. Buna göre yüksek maaş alanların memnun olduğunu, düşük maaş alanların ise işte çıktığında iş bulamama kaygısı olabileceğini düşünülebilmektedir.

Modelleme (Modeling): Bu aşamada, çalışan kaybını tahmin edebilmek için Karar ağaçları, Naive Bayes, Rasgele Ormanlar, Lojistik Regresyon ve XGBoost modelleri, önce eğitim verisi ile eğitilmiş ardından modellerin görmediği veriler üzerinden tahminleri gerçekleştirilmiştir. Veri setinde yaş, gelir, çalışma yılı vb. gibi farklı ölçeklere sahip değişkenler bulunduğundan Şekil 5'te görüleceği üzere nicel değişkenler için min-max normalizasyonu yapılmıştır. Veri seti %80 eğitim; %20 test verisi olacak şekilde bölünmüş ve ardından her bir model eğitilmiştir.



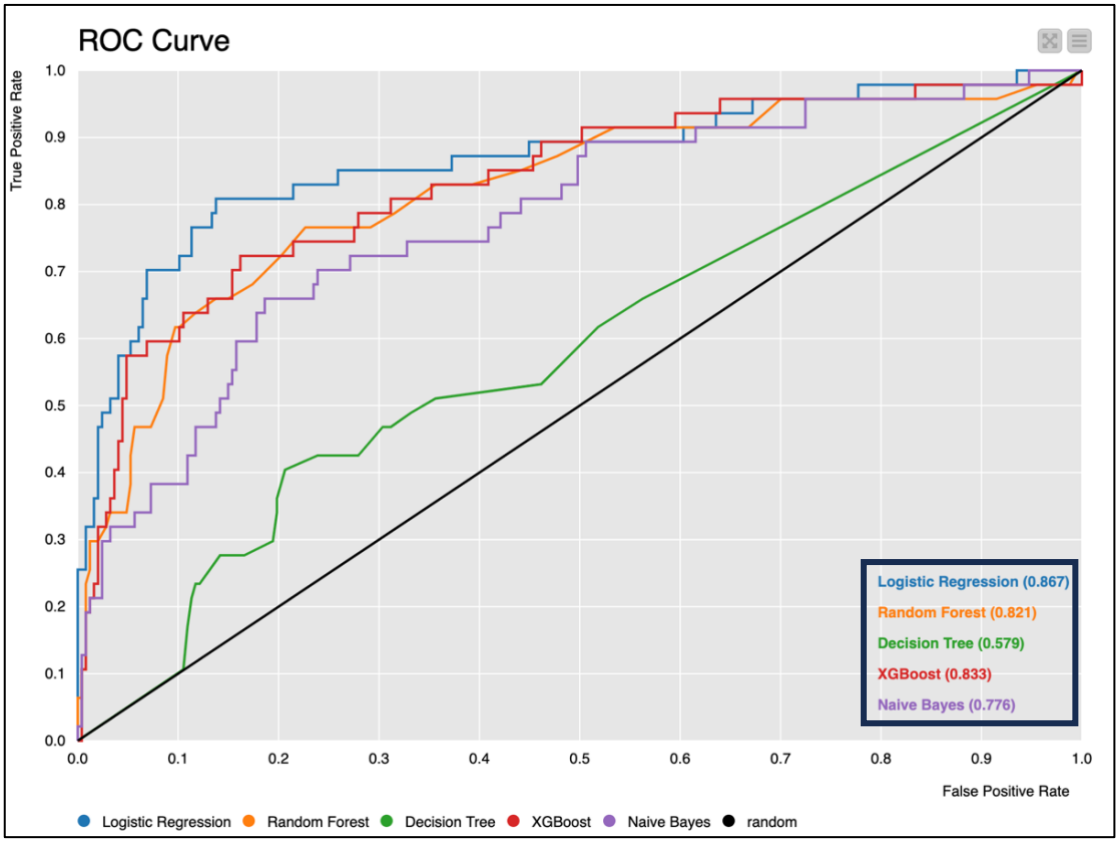
***K-katlı çapraz doğrulama yapmayı amaçlanılmış ancak model başarılarında pek bir fark yaratmadığı için ve KNIME'ın çalışma hızını yavaşlatmaması adına kaldırılmıştır.**

Şekil 5

Değerlendirme (Evaluation): Son olarak bu aşamada, oluşturulan çeşitli modellerin karşılaştırılması yapılmıştır. Modeller karşılaştırılırken doğru sınıflama oranı (accuracy) ve Cohen's Kappa başarı metrikleri kullanılmıştır. Tablo 2'de görüldüğü üzere çalışan kaybını tahmini için en yüksek oranda doğru sınıflamayı yapan model, %89,8'lik bir oran ile lojistik regresyon modeli olmuştur. Şekil 6'da görüldüğü gibi tüm modeller üzerinden oluşturulan ROC (Receiver Operator Curve) eğrisine göre logistik regresyon modelinin işten ayrılacak çalışanları tahmin etmede diğer modellere göre başarılı olduğu; karar ağaçlarının ise diğer modellere göre zayıf performans gösterdiği söylenebilmektedir.

Tablo 2

Model	Overall Accuracy	Overall Error	Cohen's Kappa	Correctly Classified	Incorrectly Classified
Decision Tree	0.75	0.252	0.109	220	74
Naive Bayes	0.765	0.235	0.337	225	69
Random Forest	0.874	0.126	0.361	257	37
Logistic Regression	**0.898**	0.102	**0.55**	264	30
XGBoost	0.871	0.129	0.393	256	38



Şekil 6

Sonuç ve Tartışma

Çalışan kaybı, şirketler için önem arz eden bir durumdur. Yüksek sayıda çalışan kaybı, şirket maliyetini arttırabilmekte, yeni bir çalışan aramak için zaman harcanmasına neden olabilmekte ve iş verimsizliği yaratabilmektedir. Bu nedenle işten çıkma olasılığı yüksek çalışanların önceden tahmini yapılarak şirketlerin müdahale etmesini sağlamak faydalı olabilecektir.

Bu amaçla yapılan çalışmada veri anlama, veri hazırlama, veri görselleştirme, modelleme, değerlendirme aşamaları tamamlanmış ve model karşılaştırma tablosuna (Tablo 2) göre lojistik regresyon modelinin %89,8'lik bir doğru sınıflandırma oranı ile diğer modellere göre daha başarılı olduğu sonucuna varılmıştır.

Çalışma kapsamında kurgusal bir veri seti yerine gerçek bir seti kullanılabilir, veri seti genişletilebilir ve daha dengeli bir veri seti oluşturulabilir.
Veri setine, kişinin iş dışı veya işteki sosyal aktiviteleri, şirketin yan imkanları (yol ücreti, yemek ücreti, sağlık sigortası vb. gibi), çalışma yeri (uzaktan-hibrit-yerinde), tatil durumu vb. gibi değişkenler dahil edilerek çalışmadan daha anlamlı sonuçlar çıkartılması sağlanabilir.
Modellerin başarısının arttılırması amacıyla ise özellik mühendisliği ve çeşitli hiperparametre optimizasyonları yapılabilir.

Kaynaklar

- **https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset (Dataset)**
- **https://www.kaggle.com/code/hamzaben/employee-churn-model-w-strategic-retention-plan**
- **https://uzmanposta.com/blog/big-data/**
- **https://www.flaticon.com/search**
- **https://hub.knime.com/**