**Hasta Kalış Süresi Tahmini**

İsa Kulaksız

**1** Atatürk Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Erzurum, Türkiye (ORCID: 0000-0000-0000-0000)

**Öz**

Hastane ortamına gelecek olan yeni vaka’nın kalma süresini yönetmek oldukça zordur.

Önceki çalışmalarda acil servis (ED) ile ilişkili çeşitli faktörlere değinilmiştir. Burada ED LOS’u daha uzun olan katılımcılar arasında daha düşük hasta memnuniyeti görülmüştür. (Parker and Marco 2014)

Bunun yanısıra Avustralya’da yoğun bakımda hastaların kalış süresi 12 farklı sütun yardımıyla tahmin edilmeye çalışılmıştır. Burada asıl amaç konvansiyonel hastaların tahmine dayalı performansını ve ve model özelliklerini karşılaştırmaktır.(Moran and Solomon 2012)

Ayrıca başka bir araştırmada ise doktorların kabul öncesi tahminlerinin, kalış süresindeki belirsizliği azaltmak ve etkili zamanlama ve insan gücü tahsisine izin vermek için yararlı bir araç olduğu sonucuna varılmıştır.(Robinson, Davis, and Leifer 1966)

Bu çalışmada ise hastanelerin optimum kaynak tahsisi ve işleyişinin her hasta bazında kalış süresinin tahmin edilmesi amaçlanmıştır.

Metod

Veri seti test ve eğitim olarak ayrı halde yer almaktadır. Burada test verisi içerisinde tahmin edilecek olan sınıfımız (Stay) olmadığı için sample\_sub.csv dosyası’nda yer alan verilerle birleştirme işlemi yapıldıktan sonra ön işleme aşamaları yapıldı: Sütunlara ait boş değerler o sütunun ortalaması ile dolduruldu ve veri seti üzerinde rahat işlem yapılabilmesi için sayısal hale dönüştürme işlemleri yapıldı.

Ardından test ve eğitim setini birleştirerek ayrı bir CSV dosyası üzerinden, test için %30 olarak ayırma işlemi yapıldı ve validasyon için ise eğitim setinin %30’u ayrıldı.

Bulgular

Test verisi üzerinde Tablo 1 de yer alan öznitelikler kullanılarak kalma süresinin (Stay) tahmin edilmeye çalışılmıştır. Bu veri seti üzerinde “Decision Tree, Random Forest, Logistic Regression, KNN, Naive Bayes” algoritmaları kullanılmıştır.

Veri setleri üzerine yeni bir öznitelik eklendi: İlk olarak burada hastanın yaşının 60 üzeri ve hastalık şiddetinin aşırı (Extreme) olması durumunda öncelik durumuna 1 verildi aksi durumda 0 verildi. Böylece veri seti ile doğrudan ilişkili olan yeni bir özellik eklenmiş oldu. Bunu yaparken ilk önce sql’de inner join yöntemi denendi fakat yapılan bu yöntem yalnızca filtre elemanlarını kapsadığı için bunun yerine outer join kullanıldı ve benzer veriler veri setinden kaldırılarak istenen bilgiler elde edildi.

Elde edilen öncelik özelliğimizi kullanarak yeniden bir karar ağaçları algoritması çalıştırıldı. Buradan çıkan sonuç: Karar ağaçları için yükselen bir doğruluk(accuracy) oranın arttığı gözlemlendi.

Sonuç

Hasta kalış süresinin tahminin’de servis tipi (Ward Type) ve hasta ile ziyaretçi sayısı (Visitor with Patient) en önemli iki özellik olarak ispatlanmış oldu.

Veri seti üzerinde öznitelik seçiminin doğruluk üzerinde ciddi bir avantaj sağlamadığı kanıtlanmış oldu.

**Hasta Kalış Süresi Tahmini**

**Abstract**

Hastanın hastanede kalış süresinin belirsizliği, etkin planlama için önemlidir. Bu belirsizliği azaltmak için teşhis bilgilerinin kullanımı gibi çeşitli yöntemler kullanılmaktadır. Hastanede doktorların hastaları kabul öncesi tahminlerinin, çeşitli özellikler yardımıyla hastanın kalış süresinde ki belirsizliği azaltmak olduğu sonucuna varılabilir.

Hasta kalış süresi çarpık dağılım gösterir ve bu tür verilerin üzerinde yapılan analizlerde çeşitli istatistiksel modelleme stratejileri kullanılmıştır.

Diagram

Description automatically generated

1. Giriş

Gün geçtikçe artan insan nüfus artışı, hastanelerin kapasitesinin yetersiz hale gelmesine neden olmuştur. Hastalar yaşadığı sağlık problemlerinde yatış işleminin yapılabilmesi ve hastanenin yatak kapasitesini etkin kullanabilmesi için bir uzman sisteme ihtitaçları doğmuştur. Bundan dolayı hastaların bazı özniteliklerini kullanarak kalma sürelerinin testpiti bir yöntem olarak görülmektedir. Dijital veri miktarının her geçen gün artmasıyla, verilerin analiz süreci önem kazanmaktadır. Bu sebeple Makine öğrenmesi algoritmaları, verileri kullanarak, makinelerin insana benzer beceri kazanmasını ve öngörüler yapmasını hedeflemektedir. Bu, Makine öğreniminin istatistiksel, veri madenciliği, örüntü tanıma gibi alanlarla yakından ilişkili olduğunu göstermektedir. Makine öğrenmesi alanının bir konusu olan gözetimli öğrenme ya da denetimli öğrenme verilen X girdi kümesinden Y çıktı kümesinin elde edilmesi için bir fonksiyon öğrenilmesidir.(Introduction to Machine Learning | The MIT Press n.d.)

Gözetimli öğrenmede, öğrenilmek istenen veri ile ilgili toplanan veri setleri bir eğitim seti olarak verilir. Eğitim kümesinde her örnek için istenen çıktı değeri de verilir. Bu bilgiler ışığında giriş ve çıkış arasında bir ilişki kurulmaya çalışılır. Oluşturulan ilişki kullanılarak gelecekte karşılaşılacak X’ gözlemlerine karşılık Y’ çıktıları tahmin edilir.

Gözetimsiz öğrenme ise verileri giriş-çıkış şeklinde etiketlemeden, veri içerisinde var olan ilişkilerin incelenmesidir.

Bu makalede amaç hastaneye gelen vakaların her vaka bazında kalış sürelerinin Makine öğrenimi algoritmaları aracılığıyla tahmin edilmesidir. Tahmin için kullanacağımız veri seti 17 öznitelik ve 1 sınıf (Stay) olarak verilmiştir. Kullanılan özniteliklerin ne işe yaradığı Tablo 2.’de gösterilmiştir.

Tahmin sınıfımız (Stay) ise 0-10 günden 100 güne kadar değişen 11 faklı sınıfa ayrılmıştır. Bu değerler Tablo 1’de gösterilmiştir.

*Tablo 1. Stay sütununda yer alan veriler*

*Chart, bar chart

Description automatically generated*

*Tablo 2. Veri setinde yer alan öznitelikler*

|  |  |
| --- | --- |
| **Öznitelikler** | **Açıklama** |
| **case\_id** | Her vaka için benzersiz kimlik |
| **Hospital\_code** | Hastane için benzersiz kod |
| **Hospital\_type\_code** | Hastane türü için benzersiz kod |
| **City\_Code\_Hospital** | City Code of the Hospital |
| **Hospital\_region\_code** | Hastanenin bulunduğu şehir kodu |
| **Available Extra Rooms in Hospital** | Hastanede mevcut ek oda sayısı |
| **Department** | Vakanın yer aldığı departman |
| **Ward\_Type** | Servis tipi |
| **Ward\_Facility\_Code** | Servis tipi kodu |
| **Bed Grade** | Servisteki yatağın durumu |
| **patientid** | Her hasta için benzersiz kod |
| **City\_Code\_Patient** | Hastanın bulunduğu yerin şehir kodu |
| **Type of Admission** | Hastanenin hastayı kabul türü |
| **Severity of Illness** | Hastalığın şiddeti |
| **Visitors with Patient** | Hasta ile ziyaretçi sayısı |
| **Age** | Hastanın yaşı |
| **Admission\_Deposit** | Giriş saatinde depozito |
| **Stay** | Hasta kalma süresi |

2. Materyal ve Metot

**2.1. Ön İşleme**

Ön işleme aşamasında boş değerler: Bed Grade 148 adet ve City\_Code\_Parient 6689 adet veri boş olarak fark edilmiştir. Burada bu iki sütunun her birinin ortalaması alınıp ilgili boş değerler üzerine yazıldı. Böylece boş değerler silinmek yerine buradan bir çıkarım elde edilebilir hale gelmiş oldu. Özniteliklerden bazıları sayısalken bazılarıysa kelime olarak verilmiş, Makine öğrenmesi algoritmalarını daha rahat uygulanabilmesi için veri setinin sayısal hale getirilmesi işlemi yapıldı.

Test ve eğitim setini birleştirdikten sonra oluşan yeni veri setinin %30’u test seti %70’i ise eğitim seti olarak ayrıldı. Burada yapay öğrenme algoritmalarından (KNN, Decision Tree, Random Forest, Logistic Regression, Naive Bayes) çıkan sonuçlar Tablo 3’te gösterilmiştir.

*Tablo 3. Doğruluk (Accuracy) oranları*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Random Forest | Naive Bayes | KNN | Logistic Regression | Decision Tree |
| %33,6 | %33,07 | %32,01 | %34,36 | %26,93 |

***2.1.1. Öznitelik Seçimi***

Veri setinde yer alan 17 adet öznitelik içerisinden Tablo 4’te yer alan öznitelikler seçilerek veri setinin tamamının üzerinde işlem yapmak yerine tahmin edilmesi gereken sınıfla doğrudan bağlantısı olmadığı düşünülen sınıflar kaldırıldı. Veri setinde yer alann yaş hastalık şiddeti ve kalma süresi bir grafik halinde Tablo 5’te gösterilmiştir. Öznitelik seçiminde sonra Makine öğrenimi algoritmaları tekrar çalıştırılarak sonuçlar Tablo6’da gösterildi.

*Tablo 4. Veri setinde yer alan öznitelikler*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Caseid | HospitalCode | Department | Age | Severity of Ilness | Type of Admission | Stay |

*Tablo 5. Yaş – Hastalık Şiddeti – Kalma Süresi*

Chart, sunburst chart

Description automatically generated

*Tablo 6. Doğruluk (Accuracy) oranları*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Random Forest | Naive Bayes | KNN | Logistic Regression | Decision Tree |
| %28,66 | %30,09 | %23,37 | %30,09 | %28,9 |

***2.1.1. Öznitelik Çıkarımı***

Veri setinde yer alan sütunları kullanarak bunlardan kalma süresi ile alakalı ayrı bir sütun oluşturuldu. Bu sütun hastanın yaşının 61 ve üzeri, hastalık şiddetinin aşırı olması durumunda öncelik değeri olarak 1 aksi durumda 0 olarak verildi. Bu sütunu oluştururken SQL’de yer alan join işlemlerinden yararlanıldı. Burada başlangıçta inner join kullanıldı fakat bu yöntem ortak kesişim değerlerini aldığı için tüm veri setinde işlem yapılmasını engelliyordu. Bundan dolayı sütunları (Age, Severity of Ilness) birleştirirken outer join yöntemi kullanıldı. Öncelik sütunuyla kalma süresi arasındaki ilişki Tablo 7’de gösterilmiştir.

*Tablo 7. Öncelik – Kalma süresi*

Chart, sunburst chart

Description automatically generated Table

Description automatically generated

Veri seti üzerinde Makine öğrenimi algoritmaları %30 test verisi olarak tekrar çalıştırıldı. Doğruluk oranları Tablo 8’de gösterilmiştir.

*Tablo 8. Doğruluk (Accuracy) oranları*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Random Forest | Naive Bayes | KNN | Logistic Regression | Decision Tree |
| %31,36 | %33,09 | %23,2 | %33,12 | %31,17 |

***2.2.0. Validasyon Oluşturma***

Validasyon bölümü için eğitim seti içerisinden %30’luk kısım ayrılmıştır. Burada validasyon bölümü için model iyileştirilmeye çalışımıştır. Bunun için hiper parametrik uygulamalar denenerek en optimum kaysayılar bulunmaya çalışıldı. Tablo 9’da validasyon için doğruluk oranları gösterilmiştir.

*Tablo 9. Doğruluk (Accuracy) oranları*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Random Forest | Naive Bayes | KNN | Logistic Regression | Decision Tree |
| %38,31 | %35,74 | %26,67 | %34,43 | %29,94 |

Hasta kalma süresini tahmini için Makine öğrenimi algoritmalarından çıkan doğruluk skorları Tablo 10’da gösterilmiştir. Burada kullanılan Makine öğrenimi algoritmalarında parametreler varsayılan olarak verildi. Örn: KNeighborsClassifier(n\_neighbors=5) gibi parametreler kullanıldı.

*Tablo 10. Doğruluk (Accuracy) oranları*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ML Algorithms | Initial Dataset | Feature Selection | Feature Extraction | Validation |
| Random Forest | %33,6 | %28,66 | %31,36 | %38,31 |
| Naive Bayes | %33,07 | %30,09 | %33,09 | %35,74 |
| KNN | %32,01 | %23,37 | %23,2 | %26,67 |
| Logistic Regression | %34,36 | %30,09 | %33,12 | %34,43 |
| Decision Tree | %26,93 | %28,9 | %31,17 | %29,94 |

Karar ağacı, bir ağaç yapısı şeklinde sınıflandırma veya regresyon modelleri oluşturur. Bir veri kümesini giderek daha küçük alt kümelere ayırırken aynı zamanda ilgili bir karar ağacı aşamalı olarak gerçekleştirir. Burada kullanılan doğruluk (accuracy) bir modelin başarısını ölçmek için kullanılan ancak tek başına yeterli olmadığı görülen bir metriktir.

Accuracy = (TP + TN) / (TP + FP + TN + FN)

Özellikle eşit dağılmayan (unbiased) veri setinde model doğruluğu tek başına yeterli değildir. Örneğin covid-19 olan ve olmayan hastaların olduğu 100 kişilik bir veri setinde tüm hastalar içinde sadece 10 tanesine covid teşhisi konulmuştur. Böyle bir durumda kanser olan ancak teşhis edilemeyen (False Negative) hastalar olması istenmez. Burada kesinlik (precision) pozitif olarak tahmin edilen değerlerin gerçekte kaç adedinin pozitif olduğunu gösterir.

Precision = TP / (TP + FP)

Bundan dolayı sıra karar ağacı algoritmasını kullanarak kesinlik (precision) ve f1 değerlerinden çıkan sonuçlar aşağıda yer almaktadır.

Kesinlik (precision): %29,51

F1: %29,51

Makine öğrenmesinde kullanılan sınıflandırma modellerinin performansını değerlendirmek için hedef niteliğe ait tahminlerin ve gerçek değerlerin karşılaştırıldığı hata matrisi sıklıkla kullanılmaktadır. Sınıflandırma tahminleri dört özelliğe sahiptir.

1. Doğruya doğru (True Positive – TP) doğru
2. Yanlışa yanlış (True Negative- TN) doğru
3. Doğruya yanlış (False Positive-FP) yanlış
4. Yanlışa doğru (False Negative-FN) yanlış

Table

Description automatically generated

3. Araştırma Sonuçları ve Tartışma

**3.1. Araştırma Sonuçları**

Hasta kalma süresinin tahmini için kullandığımız Makine öğrenimi algoritmalarında (Decision Tree, Random Forest, Logistic Regression, KNN, Naive Bayes) Tablo 10’da gösterildiği gibi öznitelik seçiminin karar ağaçları hariç bir avantaj sağlamadığı kanıtlandı. Daha önce bu veri seti üzerinde yapılan benzer çalışmalarda da benzer doğruluk oranları elde edilmiştir. Bunlardan farklı olarak burada kullanılan öznitelik çıkarımındaysa Naive Bayes ve Karar ağaçları algoritmalarından çıkan doğruluk değerlerinde bir artış olduğu kanıtlanmış oldu. Veri setinde kalma süresinin tahminin de en önemli özellikler Tablo 11’de gösterilmiştir.

*Tablo 11. En önemli özellikler*

Chart, histogram

Description automatically generated

Kalma süresinin tahmininde, hasta ile ziyaretçi sayısı’nın birbirleriyle doğrudan ilişkiye sahip olduğu Tablo 11’de gösterilmiştir.

**3.1. Tartışma**

Yapılan araştırmanın bundan önceki çalışmalardan ayırt edici farkı oluşturulan öncelik özniteliğidir. Öncelik özniteliğinde temel olarak yaşı yüksek ve hastalık durumu kritik olan hastaların kalma süresinin artacağı varsayılmıştır. Makine öğrenimi ile yapılan hesaplamalarda kullanılan algoritmaların parametreleri değiştirilerek doğruluk (accuracy) oranında bir artış gözlemlendi. Gelecekte yapılacak olan benzer çalışmalarda kullanılacak olan test veri setinde, kalma süresinin doğru sınıflandırılabilmesi için daha anlamlı bir yapıda olması gerekmektedir.

Böylece veri seti rahat bir şekilde parçalanabilir ve kalma süresi tahminindeki başarı oranı sonuçları (doğruluk, f1 vb.) daha yüksek çıkarılabilir.

4. Sonuç

Bu çalışmada hastaneye her gelen vakanın vaka bazında kalış süresi incelenmiştir. Bu inceleme hastanın hastaneye kayıt aşamasında verdiği bilgileri kapsamaktadır. Bu süreçte Makine öğrenimi algoritmalarını kullanarak hastane kapasitesinin verimli kullanılabileceği kanıtlanmıştır.

Yapılan bu çalışma sonucunda en optimum algoritmanın Random Forest algoritması olduğu Tablo 10’da gösterilmiştir. Benzer şekilde en önemli özniteliklerin ise “City Code Hospital, Type of Admission, Ward Type, Visitors with Patient” olduğu Tablo 11’de gösterilmiştir.

Çalışmada öznitelik seçiminin kalma süresi (Stay) tahmininde faydalı olduğu kanıtlandı.

Kaynakça

Kaynaklar yazılırken APA formatında yazılmasına ve bir sonraki sayfadan başlamasına dikkat ediniz.

“Introduction to Machine Learning | The MIT Press.” https://mitpress.mit.edu/books/introduction-machine-learning (April 24, 2022).

Moran, John L., and Patricia J. Solomon. 2012. “A Review of Statistical Estimators for Risk-Adjusted Length of Stay: Analysis of the Australian and New Zealand Intensive Care Adult Patient Data-Base, 2008–2009.” *BMC medical research methodology* 12(1): 1–17.

Parker, Brendan T., and Catherine Marco. 2014. “Emergency Department Length of Stay: Accuracy of Patient Estimates.” *Western Journal of Emergency Medicine* 15(2): 170.

Robinson, Gordon H., Louis E. Davis, and Richard P. Leifer. 1966. “Prediction of Hospital Length of Stay.” *Health services research* 1(3): 287.