Araştırma Yöntemleri Dersi Final Ödevi

Öykü Akın
2200329042

Hacettepe Üniversitesi
İstatistik Bölümü

15 Haziran 2023

Makalenin Yayınlandığı Dergi Hakkında Bilgiler

Derginin Adı: Journal of Data Science

Derginin İndex'i: SCI-E

Dergini Hakkında: Journal of Data Science, Çin Renmin Üniversitesi, İstatistik Okulu, Uygulamalı İstatistik Merkezi'nin resmi bir dergisidir. 2003 yılında kurulan Journal of Data Science, verilerden bilgi ve içgörülerin çıkarılacağı tüm bilimsel alanlarda veri bilimi yöntemlerini, hesaplamayı ve uygulamaları ilerletmeyi ve teşvik etmeyi amaçlamaktadır.

Derginin Erişim Linki: https://jds-online.org/journal/JDS

Makalenin İncelenmesi

Makalenin Adı: Active Data Science for Improving Clinical Risk Prediction

Makalenin Yazarları: Donna P. Ankerst ve Matthias Neumair

Makalenin Dili: İngilizce

Makalenin İncelenmesi

Makalenin Konusu: Klinik risk tahmin modellerinin geliştirilmesi ve klinik uygulama ile araştırmaya fayda sağlamak için veri bilimi yöntemlerinin kullanılması

Makalenin Amacı: Bu makale, klinik risk tahmininin daha etkili hale getirilmesi ve klinik uygulama ile araştırmaya daha fazla fayda sağlanması amacıyla veri bilimi yaklaşımlarının nasıl kullanılabileceğine odaklanmaktadır.

Makalenin İncelenmesi

Makalenin Önemi: Makale, klinik risk tahmin modellerinin geliştirilmesi ve kullanımının klinik uygulama ve araştırma alanında daha etkili hale getirilmesine odaklanmaktadır. İlgili dört veri bilimi odaklı strateji ise şu şekilde sıralanmaktadır:

- Prospektif veri toplamanması ve verilerin doğrulanması,
- Risk araçlarının ve model formüllerinin çevrimiçi olarak erişilebilir hale getirilmesi, hastaların ve klinik uzmanların nicel bilgilere daha kolay erişiminin sağlanması,
- Risk araçları ve modeller, demografik ve klinik koşullara uyum sağlamak için sürekli olarak güncellenmesi, eğitimi sırasında kohortlar arasındaki eksik veri desenleri dikkate alınması.

Makalenin Metodolojisi

Örnekleme Planı:

- Kitle: Makalede bahsedilen çalışma, çeşitli kurumlardan toplanan retrospektif prostat biyopsisi sonucu ve risk faktörü veri setlerini içermektedir. Bu veri setleri, belirli bir zaman aralığında gerçekleştirilen prostat biyopsilerine ait bilgileri içeren bir kitleden oluşmaktadır. Hangi kurumlar ve hangi zaman aralıklarıyla sınırlı olduğu net olarak belirtilmemektedir.
- Örneklem: Prostate Biopsy Collaborative Group (PBCG) kurumunun prostat biyopsisi sonuçlarını kullanmasına izin veren, 1990-2000 yılları arasında kurumda bulunmuş geçmiş hastalar.
- Örnekleme Yöntemi: Basit Rastgele Örnekleme

Makalenin Metodolojisi

Veri Toplama Yöntemi: PBCG kurumundan 1990-2000 arası geriye dönük prostat biyopsisinde risk faktörü taşıyan verilerin kurumdan alınması.

Veri Toplama Süresi: 1990 yılından başlayarak 2000'li yılların ortasına kadar veri toplandığı belirtilmiştir.

Makalenin Metodolojisi

Uygulanan Yöntemler: Çalışmada, birden fazla kurumdan alınan ileriye dönük verilere dayalı olarak prostat kanseri için bir risk tahmin aracının geliştirilmesi araştırılmaktadır. Aracı geliştirmek için çeşitli istatistiksel yöntemlerden ve makine öğrenmesinden yararlanılması amaçlanmıştır. Prostat Biyopsi İşbirliği Grubu (PBCG) kurumundan alınan veriler için, yorumlanabilirliği ve şeffaf risk tahmin formülleri nedeniyle birincil model olarak lojistik regresyon seçilmiştir. Lojistik regresyonun, öngörücü sayısı az olduğunda, rastgele ormanlar(Random Forest), K-en yakın komşular(kNN) ve yapay sinir ağları gibi makine öğrenimi yöntemlerinden daha iyi performans gösterdiği görülmüştür.

Makalenin Metodolojisi

Verilerdeki kohort etkilerini hesaba katmak için beş farklı yöntem karşılaştırılmıştır. Üç yöntem, ya kohort etkisini göz ardı ederek ya da farklı yaklaşımlar kullanarak, tüm kurumlardan gelen verileri bir havuzda toplamıştır. Diğer iki yöntem, yerel alandaki verileri analiz ederek ve tahmini regresyon katsayılarını ve bunların standart hatalarını merkezi analiz için aktararak meta-analizler gerçekleştirmiştir. Beş yöntemin tümü, doğrulamada benzer performans göstermişir.

Makalenin Metodolojisi

PBCG tarafından geliştirilen risk tahmin aracı çevrimiçi olarak kullanıma sunulmuştur ve klinisyenlere düşük dereceli ve yüksek dereceli prostat kanseri arasında ayrım yapan bir lojistik regresyon modeli sağlamıştır. Araç, altı standart risk faktörü kullanmıştır ve R Shiny uygulaması kullanılarak uygulanmıştır. Aracın genelleştirilebilirliğini değerlendirmek için dış doğrulama çalışmaları yapılmıştır ve diğer kurumları kendi risk araçlarını geliştirmeye teşvik etmek için metodolojisi ve kodu kamuoyuna açıklanmıştır. Çalışma R yazılım dili ile TRIPOD raporlama metodu kullanılarak yürütülmüştür.

Sonuç ve Tartışma

Makalede, klinik risk tahminini geliştirmek için dört veri bilimi stratejisi önerilmektedir:

İlk strateji, gelecekteki kullanımlar için verilerin aktif olarak toplanmasını ve veri giriş süreçlerinin akıllı bir şekilde tasarlanmasını vurgulamaktadır. Mevcut retrospektif verilerin temizlenmesi sürecinde harcanan kaynakların bir kısmının, geçmişte uzman olmayanlar tarafından tasarlanmış sistemlerle toplanan verilerin temizlenmesine harcandığına dikkat çekilmektedir. Bu nedenle, veri giriş süreçlerinin yeniden tasarlanması, clinician entry forms adı verilen formların değiştirilmesiyle sağlanabilir. Bu stratejinin etkili olabilmesi için her birim veya proje için zamanla uygulanması gerekmektedir.

Sonuç ve Tartışma

İkinci strateji, doğrulanmış büyük veri kohortlarına dayalı risk araçlarının çevrimiçi olarak erişilebilir hale getirilmesini önermektedir. R Shiny adlı yazılımın kullanımıyla, araştırmacılar R dilini kullanan internet tabanlı arayüzler oluşturulabilmektedir. Bu, araştırmacılar arasındaki işlemleri kolaylaştırarak zaman tasarrufu sağlamakta ve veri görselleştirme ve keşfi şeffaf hale getirmektedir.

Sonuç ve Tartışma

Üçüncü strateji, risk araçlarının düzenli olarak güncellenmesini ve yeniden uygulanmasını önermektedir. Eğer risk aracı geliştirme programları sağlıklı veri toplama süreçlerini takip ediyorsa, bu araçlar verimli bir şekilde tekrar uygulanabilmektedir. Temel yaklaşımlara yatırım yapmak, gelecekteki çalışmalarda zaman tasarrufu sağlayabilmekte ve tekrar tekrar tekerlek icat etme gereksinimini ortadan kaldırabilecek nitelikte görülmektedir.

Sonuç ve Tartışma

Son olarak, dördüncü strateji eksik verilerin hem eğitim aşamasında hem de son kullanıcı aşamasında nasıl ele alınacağını konu almaktadır. Eksik verilerin, araştırma ve yayınlarda dikkate alınması gereken önemli bir konu olduğu belirtilmektedir. Eksik verilerle başa çıkabilmek için çeşitli yöntemler bulunmaktadır ve bu yöntemlerin uygulanması için R paketleri ve öğreticiler mevcuttur. Klinik risk tahmin araçlarının kullanılabilir ve kullanıcı dostu olması için çaba harcamak gerekmektedir. Eksik risk faktörleri için seçenekler sunmak, kullanıcıya yardımcı olmak için risk faktörleri hakkında bilgi sağlamak ve hataları otomatik olarak tespit etmek için aralık kontrolleri sağlamak önemlidir.

Makale Hakkında Görüşler

Makalenin Olumlu Yönleri:

- Veri toplama sürecinin iyileştirilmesi sağlanması: Makale, veri girişi sürecinin akıllı bir şekilde tasarlanması ve veri temizliği için daha etkili yöntemlerin kullanılmasını vurgulamaktadır. Bu, gelecekteki çalışmalarda daha güvenilir ve kullanılabilir veri setlerinin elde edilmesini sağlayabilir.
- Çevrimiçi erişilebilirlik: Büyük veri kohortlarına dayalı risk araçlarının çevrimiçi olarak erişilebilir hale getirilmesi, araştırmacıların ve klinisyenlerin bu araçlardan kolaylıkla faydalanabilmesini sağlar. R Shiny gibi yazılımların kullanımı, veri görselleştirme ve keşif süreçlerini kolaylaştırarak zaman tasarrufu sağlayabilir.

Makale Hakkında Görüşler

- Güncelleme ve yeniden uygulama: Risk araçlarının düzenli olarak güncellenmesi ve yeniden uygulanması, eldeki verilerle tutarlı ve güncel sonuçlar elde etmeyi sağlamaktadır. Bu strateji, gelecekteki çalışmalarda zaman ve kaynak tasarrufu sağlayabilir.
- Eksik verilerle başa çıkma: Makale, eksik verilerin ele alınması için çeşitli yöntemlerin mevcut olduğunu ve bunların kullanılmasıyla güvenilir sonuçlar elde edilebileceğini vurgulamaktadır. Bu, araştırmacıların eksik verilerle ilgili sorunları minimize etmelerine yardımcı olabilir

Makale Hakkında Görüşler

- Kullanıcı dostu araçlar: Klinik risk tahmin araçlarının kullanıcı dostu ve kullanılabilir olması önemlidir. Makale, kullanıcıların eksik risk faktörleri için seçenekler sunulmasını ve hataları otomatik olarak tespit eden kontrollerin sağlanmasını önermektedir. Bu, araçların güvenilirliğini artırır ve kullanıcı deneyimini iyileştirebilir.
- Gelecekteki araştırmaların teşviki: Makale, yüksek boyutlu verilerle çalışan risk modellerinin daha fazla araştırmayı gerektirdiğine dikkat çekmektedir. Bu, veri bilimcilerin ve araştırmacıların yeni teknikler ve yöntemler geliştirerek klinik uygulamalarda daha iyi sonuçlar elde etmelerine yardımcı olabilir

Makale Hakkında Görüşler

Makalenin Olumsuz Yönleri:

- Zorlu uygulama süreci: Makalede önerilen bazı stratejilerin zaman ve kaynak bakımından yoğun olarak belirtilmektedir. Bu, özellikle sınırlı kaynaklara sahip olan tıbbi bilgi işlem ve istatistik merkezleri için uygulamanın zor olabileceği anlamına gelir. Bu nedenle, tüm stratejilerin tam olarak uygulanması pratik olmayabilir.
- Veri eksikliği sorunu: Makalede eksik verilerle başa çıkmanın önemi vurgulanmış olsa da, eksik verilerin ele alınması hala bir zorluk olabilir. Bu, özellikle büyük ölçekli çalışmalarda ve gerçek dünya veri setlerinde karşılaşılan bir sorundur. Eksik verilerin yanlış bir şekilde ele alınması sonuçların güvenilirliğini etkileyebilir.

Makale Hakkında Görüşler

- Yüksek boyutlu verilerin zorluğu: Makale, yüksek boyutlu verilerle çalışmanın daha fazla araştırmayı gerektirdiğini belirtmektedir. Bu tür veri setleri, analiz, işleme ve modelleme açısından zorluklar sunabilir. Yeni yöntemlerin ve tekniklerin geliştirilmesi gerekebilir.
- Teknik ayrıntıların eksikliği: Makalede önerilen stratejiler genel bir bakış sunmaktadır, ancak teknik detaylar veya pratik uygulama adımlarıyla ilgili daha fazla bilgi verilmemiştir. Bu, okuyucuların stratejileri uygulamada karşılaşacakları zorlukları anlamakta zorlanabilecekleri anlamına gelebilir.

Kaynakça

Hoogland J, van Barreveld M, Debray TPA, Reitsma JB, Verstraelen TE, Dijkgraaf MGW, et al. (2020). Handling missing predictor values when validating and applying a prediction model to new patients. Statistics in Medicine, 39(25): 3591–3607.

Jalali A, Foley RW, Maweni RM, Murphy K, Lundon DJ, Lynch T, et al. (2020). A risk calculator to inform the need for a prostate biopsy: a rapid access clinic cohort. BMC Medical Informatics and Decision Making, 20(1): 148.

Ji X, Kattan MW (2018). Tutorial: development of an online risk calculator platform. Annals of Translational Medicine, 6(3): 46.