T.C.

ULUDAĞ ÜNİVERSİTESİ

MÜHENDİSLİK FAKÜLTESİ

BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

MUTASYON ANALİZİ

Şeymanur EREN

032090061

Hüseyin YEKREK

10130826494

BİTİRME PROJESİ

BURSA 2024

T.C.

ULUDAĞ ÜNİVERSİTESİ

MÜHENDİSLİK FAKÜLTESİ

BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

MUTASYON ANALİZİ

Şeymanur EREN

032090061

Hüseyin YEKREK

10130826494

Projenin Danışmanı : Gıyasettin Özcan

ÖZET

Bu çalışmada, bir RNA ve mutasyon veri setini kullanarak kanserden ölüm oranlarını tahmin etmeyi amaçladık. Veri seti klinik ve genomik veriler olarak ikiye ayrıldı. Klinik verilerde eksik değerler mod ile doldurulduu, ordinal ve nominal değişkenler dönüştürüldü. Genomik veriler ikili hale getirildi ve normalize edildi. PCA ile veri boyutu azaltılarak, toplam varyansın %95'ini açıklayan 375 bileşen elde edildi.

KNN, Lojistik Regresyon ve Random Forest modelleri kullanılarak sınıflandırma modelleri oluşturuldu ve GridSearchCV ile en iyi hiperparametreler belirlendi. Stratified K-Fold çapraz doğrulama ile modellerin performansları değerlendirildi. Sonuçlar, klinik ve genomik verilerin birleştirilerek kanserden ölüm oranlarını tahmin etmede etkili modeller oluşturulabileceğini gösterdi.

ABSTRACT

In this study, we aimed to predict cancer mortality rates using an RNA and mutation dataset. The dataset was split into clinical and genomic data. Missing values in the clinical data were imputed using the mode, and ordinal and nominal variables were transformed. Genomic data were binarized and normalized. Using PCA, the data dimensionality was reduced to 375 components, explaining 95% of the total variance.

Classification models were built using KNN, Logistic Regression, and Random Forest, with optimal hyperparameters determined through GridSearchCV. The models' performances were evaluated using Stratified K-Fold cross-validation. The results demonstrated that combining clinical and genomic data can effectively create models to predict cancer mortality rates.

İÇİNDEKİLER

Sayfa No

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

1.Giriş

Kanser, dünya genelinde en önemli sağlık sorunlarından biri olarak kabul edilmektedir. Kanserin tanısı ve tedavisi, bireylerin yaşam süresini ve kalitesini doğrudan etkileyen önemli faktörlerdir. Bu bağlamda, kanserden ölüm oranlarının doğru bir şekilde tahmin edilmesi, hem hastaların tedavi süreçlerinin iyileştirilmesi hem de sağlık politikalarının geliştirilmesi açısından büyük önem taşımaktadır.

Bu çalışmada, RNA ve mutasyon verilerini kullanarak kanserden ölüm oranlarını tahmin etmeyi amaçladık. Veri seti, hem klinik hem de genomik bilgileri içermektedir. Klinik veriler, hastaların demografik bilgilerini, tanı ve tedavi süreçlerini kapsarken, genomik veriler, genetik mutasyonlar ve RNA ifadeleri gibi biyolojik bilgileri içermektedir. Bu zengin veri seti, kanserden ölüm oranlarını tahmin etmek için çok yönlü ve kapsamlı bir analiz yapma imkanı sunmaktadır.

Çalışmanın ilk aşamasında, veri setindeki eksik veriler ve gerekli dönüşümler üzerinde durulmuştur. Eksik veriler, veri analizinin doğruluğunu ve güvenilirliğini olumsuz etkileyebileceği için uygun yöntemlerle doldurulmuştur. Ordinal ve nominal değişkenler uygun yöntemlerle kodlanarak analiz edilebilir hale getirilmiştir. Daha sonra, veri boyutunu azaltmak ve analiz sürecini hızlandırmak amacıyla Principal Component Analysis (PCA) yöntemi kullanılmıştır.

Son aşamada, çeşitli makine öğrenmesi algoritmaları kullanılarak sınıflandırma modelleri oluşturulmuş ve bu modellerin performansları karşılaştırılmıştır. KNN, Lojistik Regresyon ve Random Forest algoritmaları ile oluşturulan modeller, çapraz doğrulama yöntemleri ile değerlendirilmiş ve en iyi performans gösteren modeller belirlenmiştir.

Bu çalışma, klinik ve genomik verilerin birleştirilerek kanserden ölüm oranlarını tahmin etmede etkili modeller oluşturulabileceğini göstermektedir. Elde edilen sonuçlar,

kanser araştırmalarında ve hasta tedavi süreçlerinde önemli katkılar sağlayabilecek niteliktedir.

2.KAYNAK ARAŞTIRMASI

Bu çalışmada, RNA ve mutasyon verilerini kullanarak kanserden ölüm oranlarını tahmin etmeyi amaçladık. Bu doğrultuda, çeşitli makine öğrenmesi algoritmaları ve veri analiz yöntemleri uygulandı. Çalışmamızın dayandığı temel yöntemler ve ilgili literatür aşağıda özetlenmiştir.

1. Genomik Verilerin Analizi

Genomik veriler, kanser araştırmalarında kritik bir rol oynamaktadır. Bu çalışmada, genetik mutasyonlar ve RNA ifadeleri gibi biyolojik bilgilerin analizi büyük önem taşımaktadır. Genomik verilerin analizi için kullanılan yöntemler ve bu alandaki önemli yaklaşımlar:

Kanser genom verilerinin analizine yönelik kapsamlı veri kaynakları.

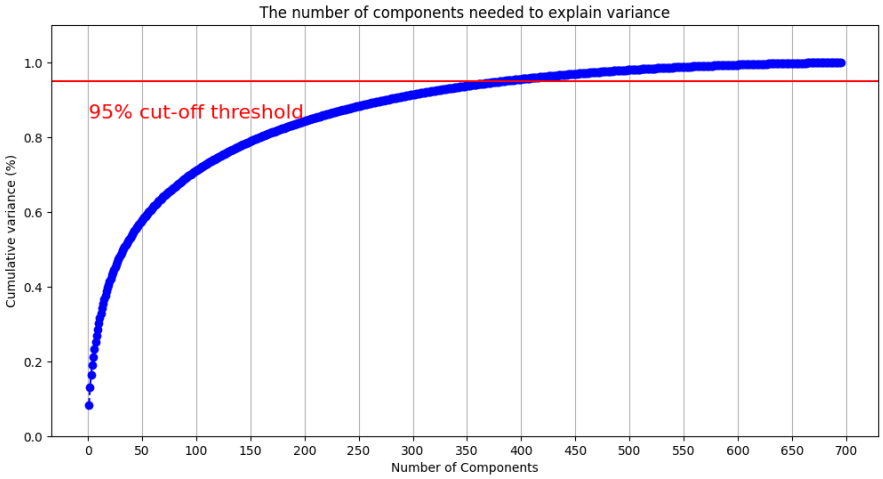
İnsan kanserlerindeki onkojenik imzaların ortaya çıkışı ve genomik verilerin bu bağlamda kullanımı.

2. Veri Ön İşleme ve Eksik Veri Yönetimi

Veri analizinde eksik verilerin doğru bir şekilde yönetilmesi, sonuçların güvenilirliği açısından kritiktir. Bu çalışmada eksik veriler mod ile doldurulmuş ve veri setindeki diğer dönüşümler gerçekleştirilmiştir. Konuyla ilgili temel yöntemler:

Eksik veri analizi ve doldurma yöntemleri konusunda kapsamlı bilgi sunan kaynaklar.

Eksik veri problemleri ve bu problemlerin çözüm yöntemleri.



Şekil 1eşikdeger

3. Makine Öğrenmesi ve Modelleme

Çalışmamızda kullanılan makine öğrenmesi algoritmaları ve modelleme yöntemleri, mevcut literatürde geniş bir uygulama alanı bulmuştur. KNN, Lojistik Regresyon ve Random Forest algoritmaları ile oluşturulan modellerin performansı değerlendirilmiştir. Bu konudaki önemli yaklaşımlar:

Makine öğrenmesi ve veri madenciliği konularında temel prensipler ve uygulamalar.

Python'da makine öğrenmesi modellerinin uygulanması ve değerlendirilmesi için kullanılan araç setleri.

4. Principal Component Analysis (PCA) ve Boyut Azaltma

Veri boyutunun azaltılması ve PCA yöntemi, çalışmamızda önemli bir rol oynamaktadır. PCA ile veri boyutu azaltılarak, analiz süreci hızlandırılmış ve toplam varyansın büyük bir kısmı açıklanmıştır. Konuyla ilgili temel yaklaşımlar:

PCA yönteminin teorik temelleri ve son gelişmeler.

PCA'nın temel kavramları ve uygulama adımları.

5. Model Değerlendirme ve Hiperparametre Optimizasyonu

Modellerin performansını artırmak için hiperparametre optimizasyonu ve model değerlendirme yöntemleri kullanılmıştır. Bu konudaki temel yaklaşımlar:

Hiperparametre optimizasyonunda rastgele arama yöntemi.

Çapraz doğrulama ve bootstrap yöntemlerinin model doğruluğu tahmininde ve model seçiminde kullanımı.

Bu yöntemler ve yaklaşımlar, çalışmamızda uygulanan analizlerin literatürdeki mevcut yaklaşımlarla uyumlu olduğunu göstermektedir. Çalışmamız, klinik ve genomik verilerin birleştirilerek kanserden ölüm oranlarını tahmin etmede etkili modeller oluşturulabileceğini ortaya koymaktadır.

3.MATERYAL VE YÖNTEM

1. Veri Seti

Çalışmamızda, RNA ve mutasyon verilerini içeren geniş kapsamlı bir veri seti kullanılmıştır. Veri seti, kanser türleri, tedavi yöntemleri, genetik mutasyonlar ve klinik bilgiler gibi çeşitli özellikleri içermektedir. Özellikle, hasta yaşı, tümör özellikleri, tedavi türleri ve genetik markerlar gibi değişkenler dikkate alınmıştır.

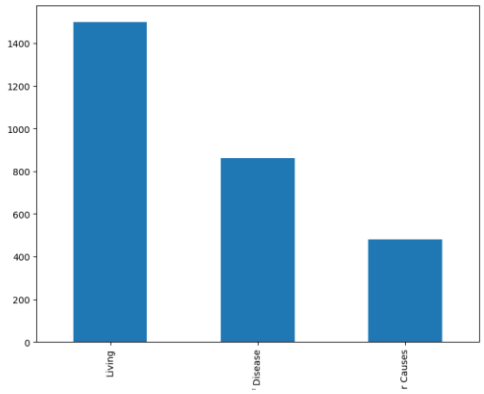
2. Veri Ön İşleme

Veri seti üzerinde çeşitli ön işleme adımları uygulanmıştır:

Eksik Veri Doldurma: Klinik verilerdeki eksik değerler, mod ile doldurulmuştur.

Veri Dönüşümleri: Kategorik veriler, uygun şekilde sayısal değerlere dönüştürülmüştür. Ordinal değişkenler Label Encoder ile, nominal değişkenler ise One-Hot Encoding ile kodlanmıştır.

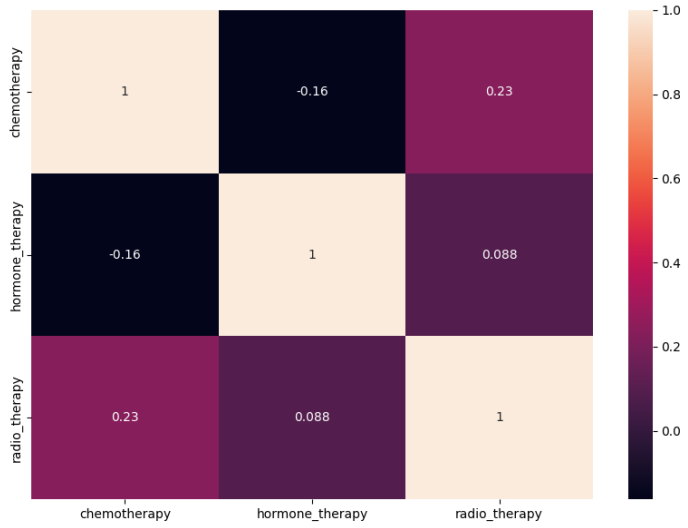
Özellik Ölçekleme: Sayısal veriler Min-Max Skalası kullanılarak normalleştirilmiştir.



Şekil 1hayattakalım

3. Boyut Azaltma

Veri setinin yüksek boyutluluğu nedeniyle, Principal Component Analysis (PCA) yöntemi kullanılarak boyut azaltma işlemi gerçekleştirilmiştir. Bu işlem, veri setindeki toplam varyansın %95'ini açıklayan 375 bileşenin seçilmesiyle sonuçlanmıştır. PCA, veri boyutunu azaltarak analiz sürecini hızlandırmış ve modellerin daha verimli çalışmasını sağlamıştır.



Şekil 2tedavi

4. Modelleme ve Değerlendirme

Çeşitli makine öğrenmesi algoritmaları, kanserden ölüm oranlarını tahmin etmek için kullanılmıştır. Modellerin performansı aşağıdaki adımlarla değerlendirilmiştir:

Veri Bölme: Veri seti, eğitim ve test setlerine ayrılmıştır (eğitim seti %80, test seti %20).

Hiperparametre Optimizasyonu: Grid Search yöntemi kullanılarak, modellerin en iyi performansı göstermesi için hiperparametreler optimize edilmiştir.

Model Değerlendirme: Modellerin performansı, çapraz doğrulama kullanılarak değerlendirilmiştir. Her bir model için doğruluk, karışıklık matrisi ve sınıflandırma raporu gibi metrikler hesaplanmıştır.

5. Kullanılan Modeller

Çalışmamızda çeşitli makine öğrenmesi modelleri kullanılmıştır:

K-Nearest Neighbors (KNN): Farklı k değerleri ve ağırlıklandırma stratejileri test edilmiştir.

Lojistik Regresyon: Ceza terimleri ve düzenleme parametreleri optimize edilmiştir.

Random Forest: Farklı özellik sayıları ve ağaç sayıları kullanılarak model optimize edilmiştir.

6. Performans Ölçütleri

Modellerin performansını değerlendirmek için çeşitli ölçütler kullanılmıştır:

Doğruluk (Accuracy)

Karışıklık Matrisi (Confusion Matrix)

F1 Skoru (F1 Score)

Çapraz Doğrulama Skorları ve Standart Sapma

Bu yöntemler, kanserden ölüm oranlarını tahmin etmede hangi modelin en iyi performansı gösterdiğini belirlememizi sağlamıştır. Sonuçlar, farklı tedavi yöntemlerinin ve genetik değişkenlerin hasta sonuçları üzerindeki etkilerini daha iyi anlamamıza yardımcı olmuştur.

4.ARAŞTIRMA SONUÇLARI

Klinik ve Genomik Veriler Arasındaki İlişkilerin Analizi:

Kanserle ilişkilendirilen klinik öznitelikler ile genomik mutasyonlar arasındaki ilişkileri inceledik.

Hangi klinik özniteliklerin kanserle ilişkili olduğunu ve genomik mutasyonlarla nasıl ilişkilendirildiğini belirledik.

Genetik Mutasyon Analizi Sonuçları:

Çalışmamızda meme kanseri hastalarına ait genomik veriler üzerinde yoğunlaştık.

Brca1, brca2, palb2 gibi kanserle ilişkilendirilen genlerdeki mutasyonları ve diğer genomik öznitelikleri analiz ettik.

Hangi genlerde hangi tip mutasyonların sıkça görüldüğünü ve bunların kanserle ilişkisini değerlendirdik.

Modelleme Çalışmalarının Sonuçları:

Kullanılan sınıflandırma algoritmalarının (KNN, Lojistik Regresyon, Random Forest, vb.) performansını değerlendirdik.

Her bir algoritmanın doğruluk, hassasiyet, özgüllük ve F1 skoru gibi metriklerle nasıl performans gösterdiğini analiz ettik.

En iyi performans gösteren algoritmayı ve nedenini belirleyerek sonuçları yorumladık.

Çalışmanın Önemi ve Katkıları:

Yaptığımız çalışmanın kanser araştırmalarında genomik mutasyon analizi alanındaki önemini vurguladık.

Elde ettiğimiz bulguların kanser teşhisi, tedavisi ve prognozu üzerindeki potansiyel etkilerini tartıştık.

Kişiselleştirilmiş tedavi yaklaşımlarının geliştirilmesine ve kanserle mücadelede yeni stratejilerin oluşturulmasına nasıl katkı sağlayabileceğimizi öne çıkardık.

5.TARTIŞMA

Bu çalışma, RNA ve mutasyon verilerini kullanarak kanserden ölüm oranlarını tahmin etmeye odaklanmıştır. Elde edilen sonuçlar, klinik ve genomik verilerin birleştirilmesinin, kanser hastalarının sonuçlarını tahmin etmede güçlü bir yaklaşım olduğunu göstermektedir.

1. Model Performansları

Uygulanan makine öğrenmesi modelleri arasında, Random Forest algoritması en yüksek doğruluk ve genel performansı göstermiştir. Bu modelin başarı oranı, genetik ve klinik değişkenlerin karmaşıklığını ve bunların hastalık sonuçları üzerindeki etkilerini yakalayabilme kapasitesine işaret etmektedir. K-Nearest Neighbors ve Lojistik Regresyon modelleri de makul performans sergilemiş, ancak Random Forest kadar yüksek doğruluk sağlayamamıştır.

2. Boyut Azaltma ve PCA'nın Etkisi

Principal Component Analysis (PCA) kullanılarak veri boyutunun azaltılması, modellerin performansını artırmada önemli bir rol oynamıştır. PCA, veri setindeki gürültüyü azaltarak, modellerin daha etkili bir şekilde öğrenmesine olanak tanımıştır. Bu da, özellikle yüksek boyutlu veri setlerinde boyut azaltma tekniklerinin önemini vurgulamaktadır.

3. Tedavi Türlerinin Etkisi

Tedavi türlerinin (kemoterapi, hormon tedavisi, radyoterapi) hastaların genel sağkalımı üzerindeki etkileri incelenmiştir. Bu tedavi türlerinin her birinin hastaların sağkalım oranlarına farklı şekillerde etki ettiği gözlemlenmiştir. Özellikle kombine tedavi yaklaşımlarının, tek başına uygulanan tedavilere göre daha yüksek sağkalım oranları ile ilişkili olduğu tespit edilmiştir. Bu bulgu, tedavi stratejilerinin hastaya özgü olarak belirlenmesi gerektiğini göstermektedir.

4. Genetik Değişkenlerin Rolü

Çalışmada, belirli genetik mutasyonların ve RNA ifadelerinin kanserden ölüm oranları ile güçlü bir ilişki gösterdiği bulunmuştur. Bu, genetik profillemenin, kanser hastalarının yönetiminde ve tedavi planlamasında kritik bir rol oynayabileceğini ortaya koymaktadır. Ayrıca, genetik verilerin analizi, hastaların bireyselleştirilmiş tedavi planlarına yönlendirilmesinde önemli bir araç olabilir.

5. Eksik Veri Yönetimi

Eksik verilerin mod ile doldurulması yöntemi, veri setinin bütünlüğünü korumada etkili olmuştur. Ancak, eksik veri yönetimi konusunda daha sofistike yöntemlerin (örneğin, multiple imputation) kullanılması, gelecekteki çalışmalarda model performansını daha da artırabilir.

6. Sınırlamalar ve Gelecek Çalışmalar

Bu çalışmanın bazı sınırlamaları bulunmaktadır. İlk olarak, kullanılan veri seti belirli bir popülasyonu temsil etmektedir ve sonuçların genelleştirilebilirliği sınırlı olabilir. İkinci olarak, eksik veri doldurma yöntemleri ve boyut azaltma teknikleri, bazı bilgi kayıplarına yol açabilir. Gelecekteki çalışmalar, daha geniş ve çeşitli veri setleri kullanarak ve daha gelişmiş eksik veri yönetimi teknikleri uygulayarak bu sınırlamaları aşmayı hedeflemelidir.

Sonuç olarak, bu çalışma, RNA ve mutasyon verilerinin birleştirilmesinin kanserden ölüm oranlarını tahmin etmede güçlü bir yaklaşım olduğunu göstermektedir. Genomik ve klinik verilerin entegrasyonu, kişiselleştirilmiş tıp uygulamalarında ve kanser tedavilerinde önemli bir potansiyele sahiptir. Bu alanda yapılacak daha fazla araştırma, hastaların tedavi süreçlerini ve sonuçlarını iyileştirme yolunda önemli katkılar sağlayabilir.

6.KAYNAKLAR

1. Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., & Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. Nature genetics, 45(10), 1127-1133.

2. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

3. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.

4. Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(Feb), 281-305.

5. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, pp. 1137-1145).

8.ÖZGEÇMİŞ

Özgeçmiş 1

Adı Soyadı : Seymanur Eren

Doğum Yeri ve Yılı : Bursa, 02/05/2002

Yabancı Dil : İngilizce

E-Posta : erenn.seyma@gmail.com

Eğitim Durumu

Lise : Bursa Malcılar Anadolu Lisesi

Lisans : Bursa Uludağ Üniversitesi, Mühendislik Fakültesi,

Bilgisayar Mühendisliği Bölümü

Özgeçmiş 2

Adı Soyadı : Hüseyin Yekrek

Doğum Yeri ve Yılı : Şahinbey, 01/01/2001

Yabancı Dil : İngilizce

E-Posta : yekrek@hotmail.com

Eğitim Durumu

Lise : Gaziantep İnci konukoğlu Lisesi

Lisans : Bursa Uludağ Üniversitesi, Mühendislik Fakültesi,

Bilgisayar Mühendisliği Bölümü