Rahim Ağzı Kanseri Veri Kümesi ile Makine Öğrenmesi Model Tanımlanabilirliği

YUSUF FURKAN YÜCESOY
152120151005

1) Giriş

1.1 Makine Öğrenmesi Model Tanımlanabilirliği Nedir ?

Günümüzde, bir kişinin kredisini geri ödeyip ödeyemeyeceğini veya belirli hastalıkları olup olmadığı gibi etmenleri tahmin etmek için makine öğrenmesinden faydalanılmaktadır. Modern dünyada makine öğrenmesinin kullanımına dair sonsuz örnek verebiliriz. Ancak, makine öğrenmesinin avantajları olduğu gibi, bazı dejavantajlarla da karşı karşıyayız. En büyük dezavantajlarından biri 'makine öğrenmesi modelinin tanımlanabilirliği eksikliği' dir.

Farz edelim ki, bir e-ticaret sitesindeki bir ürünün fiyatını tahmin eden bir modelimiz var. Bazı verileri girdi olarak alır ve sonucu (fiyat) verir. Fakat modelin neden bu tür fiyat ürettiğini biliyor musunuz? Başka bir senaryo ise; hastanın belirli bir kanser türüne sahip olup olmadığını ön gören bir modelimiz var, model çıkışı yanlış olduğu durumda ciddi hasara ve soruna yol açabilir. Böyle bir durumda sadece modele güvenebileceğimizi söyleyemeyiz. Modelin neden belli değeri çıktı/tahmin bilmemiz gerekir. Buna makine öğrenmesi modeli tanımlanabilirliği denir.

1.2 Model Tanımlanabilirliğinin Önemi ?

- Makine öğrenmesi hayatın bir parçası olduğu için modelimizin neden böyle bir öngörüde bulunduğunu bilmek isteriz. Bu, modelin güvenilirliği ve genel kullanım durumunu oluşturacaktır.
- Modelimiz, eğitim verilerinden önyargıları (bias) seçebilir. Bu makine öğrenmesi modelimizi bazı gruplara karşı ayrımcılık yapan ırkçı bir modele dönüştürebilir. Model tanımlanabilirliğini açıklayarak önyargı (bias) problemini daha iyi çözebiliriz.
- Makine öğrenmesi modelleri güvenlik önlemleri ve testler gerektiren gerçek dünyadaki görevleri üstlenir. Otonom araçlar gibi modellenlerin kullanıldığı sistemlerde barikatların, araçların ve yayaların dikkatlice tanımlandıklarından emin olmak istenilmektedir. Tek bir hatanın büyük hasara yol açabileceği bu tür senaryolarda model tanımlanabilirliği bize yardımcı olur.

1.3 Rahim Ağzı Kanseri Nedir?

Rahim ağzı kanseri yılda yaklaşık 500.000 kişiyi etkileyen, kadınlarda görülen 4. en sık kanserdir. Ayrıca, Dünya Sağlık Örgütü raporuna göre, rahim ağzı kanseri gelişmekte olan

ülkelerdeki kadınlar arasındaki en yaygın kanserdir. Genellikle 50 yaş civarında ortaya çıkan rahim ağzı kanseri son yıllarda genç kadınlarda da görülmeye başlanmıştır.

2) Proje Hakkında

2.1 Proje Amacı

Projenin amacı rahim ağzı kanseri gibi hayati önem taşıyan bir mesele üzerinde dengesiz veri kümesi ile çalışarak rahim ağzı kanseri öngörmede tıbbi olarak kullanılan modelin tanımlanabilirliğini sağlamaktır.

- Çalışma esnasında, random forest makine öğrenmesi modeli üzerinde test yapılacaktır.
- Kısmı bağımlılık grafikleri (Partial Dependence Plots) ile her bir özelliğin başarıma etkisi belirlenecektir.
- Özelliklerin model için önem analizinde "Permutasyon Önemi" fonksiyonu kullanılacaktır.
- SHAP değerleri (Shapley Additive exPlanations'ın kısaltması), belirli bir özelliğin öngörümüzü ne kadar değiştirdiği üzerinde durulacaktır.

Böylelikle rahim ağzı kanseri gibi riskli bir durumun öngörüsünde kullanılan özelliklerin hastalık tespiti üzerinde kullanılabilirliği analiz edilecektir.

Çalışma esnasında; Anaconda programını kullanarak, Python3 dili üzerindeki veri madenciliği ve makine öğrenmesi kütüphaneleriden ve modelleriden (Pandas, Numpy, Sklearn, Shap, Imblearn, Eli5 vb.) faydalanılacaktır.

2.2 Projede Kullanılacak Veri Kümesi

Kullanacak veri kümesi UCI Machine Learning Deposundan temin edilmiş Cervical Cancer (Risk Factors) Veri Kümesidir. Veri kümesi, 2017 yılında 858 örnek ve 4'ü hedef olmak üzere 36 özellik olarak yayınlanmıştır. Bu özellikler demografik bilgileri, sigara içme alışkanlıklarını ve tarihi tıbbi kayıtları içermektedir. Bu verilerin karmaşıklığı, karmaşık bir ekosisteme yol açan çoklu tarama ve teşhis yakaşımlarıdır. Sonuç olarak, hastanın faktör riskinin öngörülmesi ve en iyi tarama stratejisi temel bir sorundur. Çalışma esnasında veri kümesi veri ön işleme tekniklerine tabii tutulacaktır.

<u>Veri ön işleme neden yapılır?</u>
Veri setlerinin bilgi içeriklerinin veri madenciliği araçları tarafından en iyi şekilde analiz edilmesini sağlamak için ön işleme yapılır. Bu sayede önceden yapılan hatalı tahmin oranlarının azalması beklenir. Bu işlemler aşağıda belirtilen en az birisi varsa yapılması gerekir.
Eksik (incomplete) Veri : Toplanamamış verilerin sebep olduğu sorundur.
Gürültülü (noisy) Veri : Gürültü orijinal değerlerin bozulması anlamına gelir.
Tutarsız (inconsistent) Veri : Nitelik değerleri veya nitelik isimlerinin uyumsuz olması sorunudur.

Veri ön işleme aşamaları:

Veri Temizleme (data cleaning): Eksik değerleri doldurma; gürültülü, kirli ve pürüzlü veriyi pürüzsüz hale getirme(yumuşartma: smoothing); aykırı gözlemleri belirleme veya ayıklama ve tutarsızlıkları giderme işlemleridir.
Veri Entegrasyonu (data integration): Birden fazla veritabanının, veri küpünün (data cube), tablonun veya dosyanın entegrasyonu işlemidir.
Veri Dönüşümü (data transformation): Normalizasyon ve yığınlama (aggregation) işlemleridir. Bu aşamada veri standart hale getirilmeye çalışılır.
Veri Azaltma (data reduction): Veriyi orijinal halini temsil edecek ve aynı veya benzer analitik sonuçları üretecek şekilde hacimsel olarak küçültme, azaltma işlemleridir. Azaltılan veri, normal veriyi temsil etme özelliğini korumalıdır.
Veri Ayrıştırma (data discretization): Veri azaltma (data reduction) işleminin bir parçasıdır ve özellikle numerik verilerde dikkatle uygulanmalıdır.

	Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	Hormonal Contraceptives	Hormonal Contraceptives (years)
0	18	4.0	15.0	1.0	0.0	0.0	0.0	0.0	0.0
1	15	1.0	14.0	1.0	0.0	0.0	0.0	0.0	0.0
2	34	1.0	?	1.0	0.0	0.0	0.0	0.0	0.0
3	52	5.0	16.0	4.0	1.0	37.0	37.0	1.0	3.0
4	46	3.0	21.0	4.0	0.0	0.0	0.0	1.0	15.0
5	42	3.0	23.0	2.0	0.0	0.0	0.0	0.0	0.0
6	51	3.0	17.0	6.0	1.0	34.0	3.4	0.0	0.0
7	26	1.0	26.0	3.0	0.0	0.0	0.0	1.0	2.0
8	45	1.0	20.0	5.0	0.0	0.0	0.0	0.0	0.0
9	44	3.0	15.0	?	1.0	1.266972909	2.8	0.0	0.0

Tablo 1 - Veri setinin bir kısmı

Veri setindeki öznitelikler:

Age : Yaş
Number of sexual partners : Birlikte olunan partner sayısı
First sexual intercourse(age) : İlk birliktelik yaşı
Number of pregnancies : Hamilelik sayısı
Smokes : Sigara içip içilmeme durumu
Smokes (years) : Yıllık içilen sigara
Smokes (packs/year) : Yıllık tüketilen sigara paketi

Hormonal Contraceptives(years) : Yıllık tüketilen hormonal gebelik önleyici
IUD : Doğum kontrol cihazı kullanımı
IUD (years) : Yıllık kullanılan doğum kontrol cihazı kullanımı
STDs : Cinsel yolla bulaşan hastalık
STDs (number) : Cinsel yolla bulaşan hastalık sayısı
STDs (condylomatosis) : Hastalık türü
STDs (cervical condylomatosis) : Hastalık türü
STDs:vaginal condylomatosis : Hastalık türü
STDs:vulvo-perineal condylomatosis : Hastalık türü
STDs:syphilis : Hastalık türü
STDs:pelvic inflammatory disease : Hastalık türü
STDs:genital herpes :Hastalık türü
STDs:molluscum contagiosum : Hastalık türü
STDs:AIDS : Hastalık türü
STDs:HIV : Hastalık türü
STDs:Hepatitis B : Hastalık türü
STDs:HPV : Hastalık türü
STDs: Number of diagnosis : Teşhis edilen hastalık sayısı
STDs: Time since first diagnosis : İlk tanıdan bu yana geçen süre
STDs: Time since last diagnosis : Son tanıdan bu yana geçen süre
Dx:Cancer : Daha önce rahim ağzı kanseri tanısı konulmuş
Dx:CIN : Daha önce servikal intrapitelyal neoplazi tanısı konulmuş
Dx:HPV : Daha önce human papilloma virus tanısı konulmuş
Dx:

Hinselmann: target variable
Schiller: target variable
Cytology: target variable
Biopsy: target variable : Makale taramaları sonucu projede başarım için kullanılacak özellik

2.3 Problemi Çözümü için Kullanılabilecek Yöntemler

Dengesiz veri için:

Önyargılı verileri ele almak için çeşitli yöntemler vardır. Bu çalışma, Python dilinde imblearn sınıfında mevcut olan SMOTETomek (birleştirme yöntemi), örneklemenin az olması ve örneklemenin üstesinden gelme yöntemini uygulanacaktır.

En önemli özellikleri bulmak için:

Rahim ağzı kanseri modelimiz birçok özellik kullanıyor. Ancak rahim ağzı kanserini öngörmede hangi özelliklerin çok önemli olduğunu ve hangilerinin daha az önemli olduğunu nasıl biliyor olmamız gerekli. Burada "Özelliğin Önemi" tanımında faydalanılacaktır.

Özelliğin önemi çok klasik ve popüler bir yöntemdir. Tanıma göre, özellik önemi "Özelliğin önemi; özellik ile gerçek sonuç arasındaki ilişkiyi kıran özellik değerlerine izin verdikten sonra, modelin tahmin hatasının artmasıdır."

Özelliğin önemi, model yerleştirildikten sonra hesaplanır. Özellikler sütunlarında, her bir sütunu hedef ve diğer özellik sütununa dokunmadan rasgele karıştırır ve yeni karıştırılmış veri tahmininin doğruluğunu nasıl etkilediğini hesaplar.

Bir özellik sütununun değerinin değiştirilmesi modelin doğruluğunu yoğun şekilde etkilerse, o sütun özelliğinin önemi daha yüksektir. Bu sayede özellik önemi hesaplanır.

Özelliğin önemini hesaplamanın birçok yolu vardır. Bu defterde, eli5 kütüphanesinde 'Permütasyon önemi' metodunu kullanılacaktır.

Tek bir özelliğin tahmin üzerindeki etkisinin analizi:

Tek özelliklerin tahminimizi nasıl etkilediğini bilmek için "Kısmi Bağımlılık Grafikleri" adı verilen farklı bir teknik kullanmamız gerekir.

Kısmi Bağımlılık Grafikleri veya PDP de çok popüler bir yöntemdir. PDP, model takıldıktan sonra hesaplanır. Daha sonra sonucu tahmin etmek için test verilerinden tek bir satır kullanılır. Bir öngörmeyi tahmin etmek yerine, bir dizi tahmin yapmak için art arda satırın bir değişkenini değiştirilir. Örneğin, rahim ağzı kanseri modelimiz için test verilerinden bir satır alınır ve tekrar tekrar yaş gibi tek bir değişken değeri değiştirilir ve ardından bir dizi tahmin değeri elde edilir. Bunları çoklu satırlar için tekrarlandıktan sonra dikey eksene çıktıların ortalamaları bastırılır.

Bu projede Kısmi Bağımlılık Grafikleri için PDPBox kütüphanesi kullanılacaktır.

SHAP değerleri:

SHAP'ın amacı, bir x örneğinin öngörüsünü, her özelliğin öngörüme katkısını hesaplayarak açıklamaktır. SHAP açıklama yöntemi Shapley'nin koalisyon oyunu teorisinden elde ettiği değerleri hesaplar. Bir veri örneğinin özellik değerleri bir koalisyondaki oyuncular olarak hareket eder. Shapley değerleri tahminleri özellikler arasında nasıl adil bir şekilde dağıtacağımızı anlatıyor.

2.3 Kullanılan Veri Kümesi ile Daha Önce Yapılmış Çalışmaların İncelenmesi

- **Classification of Cervical Cancer Dataset- Y. M. S. Al-Wesabi, Avishek Choudhury, Daehan Won Binghamton University, USA**
  Çalışmada rahim ağzı kanseri veri kümesi üzerinde sınıflandırma çalışmaları yapılmıştır. Özetlemek gerekirse, bu makale farklı makine öğrenme sınıflayıcıları arasında en iyisine göre karşılaştırmalar sunmaktadır. . Sonuçlar bu verinin önyargılı olduğunu ve dengesiz verilerin

ele alınmasının değerlendirme için ilk adım olduğunu göstermektedir. Dengesiz veriyi ele almak için üç teknik kullanılmıştır; aşırı örnekleme, düşük örnekleme ve her iki yöntemi de birleştirir. Aşırı örnekleme, aşırı örneklemeyle elde edilen daha yüksek doğruluk nedeniyle diğer iki yöntemden daha iyi sonuçlar vermektedir.

- **Transfer Learning with Partial Observability Applied to Cervical Cancer Screening- Kelwin Fernandes , Jaime S. Cardoso, and Jessica Fernandes, Universidad Central de Venezuela, Caracas, Venezuela**
  Bu çalışmada rahim ağzı kanseri veri kümesi ön hastalık tanısı için örneklem olarak kullanılmıştır. Yapılan çalışma görüntüleme üzerinedir.

- **Data-Driven Diagnosis of Cervical Cancer With Support Vector Machine-Based Approaches - WEN WU AND HAO ZHOU**
  Çalışamada bazı rahim ağzı kanseri risk faktörleri gözden geçirilmiştir ve rahim ağzı kanseri veri setinin sınıflandırılmasında SVM tabanlı yaklaşım uygulanmaktadır. Standart SVM yöntemi, malign kanseri ve iyi huylu kanseri iyi sınıflandırır. Hem SVM-RFE hem de SVM-PCA, benzer işlevi SVM'den daha az özellikle gerçekleştirebilir. Daha spesifik olarak, SVM-RFE ve SVM-PCA, sınıflandırmayı gerçekleştirmek için özellik numaralarını 30'dan 8'e düşürme özelliğine sahiptir. Bu arada, sınıflandırma hızı belirgin bir şekilde iyileştirilebilir. Ayrıca, SVM yöntemi rahim ağzı kanseri verilerini kesin olarak sınıflandırabilmesine rağmen, yüksek hesaplama maliyeti bir sınırlama olarak görülmektedir.

3)Deney

Deneyde amaç rahim ağzı kanseri gibi riskli konular üzerine geliştirilen modellerde başarım oranı parametresinin yeterli güvene sahip bir parametre olmadığını ortaya koyup, farklı ölçüm parametreleri ve tekniklerle bu yetersizliği ortaya koymak ve geliştirilen modele tanımlanabilirlik sağlamaktır.

3.1 Veri Seti Üzerinde Oynama yapılmaksızın Random Forest Sınıflandırıcı Algoritmasındaki Başarımının Testi

Yapılan literatür taraması sonucu hedef öznitelik olarak ‘Biopsy’ seçilmiş ve özniteliğin adı çalışmada karışıklılığa yol açmamak adına ‘Cancer’ olarak değiştirilmiştir.

Random Forest Sınıflandırıcı ile çalışmak adına veri setimiz %25-%75 olacak şekilde test ve train olarak ikiye ayrılmıştır. Yapılan eğitim ve test sonucunda elde elidlen başarım skoru tabloda yer almaktadır.

Accuracy score:	0.9627906976744186

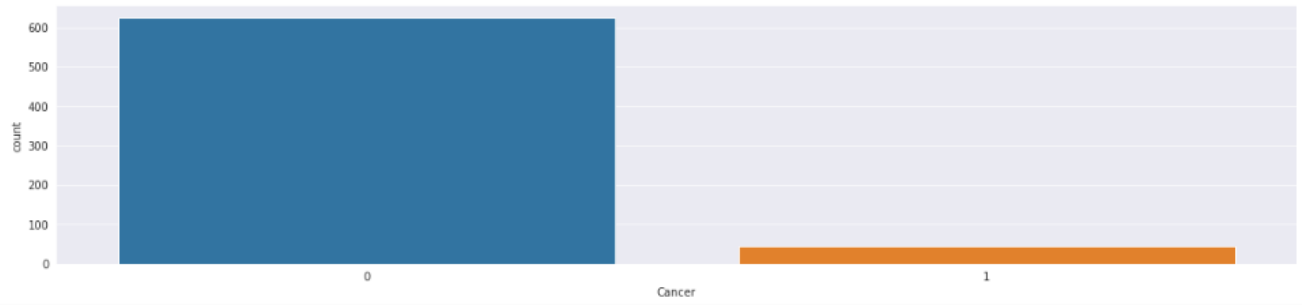
Tablo 2 - Başarım Skoru

Hiçbir ön işleme tabii tutulmamış içinde eksik değerler bulunan ayrıca dengesiz bir dağılım gösteren veri setimiz %95'in başarım skoru vermektedir. Fakat şuan modelimizin bir hastayı kanser veya kanser değil olarak tahminini neye göre yaptığı konusunda fikrimiz yoktur. Literatürde bu durum 'black box' (kara kutu) problemi olarak adlandırılmaktadır. Yaygın olarak benimsenmesine rağmen, makine öğrenmesi modelleri çoğunlukla kara kutu olarak kalmaktadır. Günümüzde modeller başarım skoru kullanılarak değerlendirilmektedir. Başarım skorunun modele olan güveni ne derece sağlayabilecek bir parametre olabileceğini test etmek adına veri setimizden rastgele seçilmiş (UID) bir öznitelik ve hedef değer olan Cancer özniteliğini %25-%75 olarak test ve train olarak ayırılmıştır. Yapılan eğitim ve test sonucunda elde edilen başarım skoru tabloda yer almaktadır.

Accuracy score:	0.9488372093023256

Tablo 3 - Başarım Skoru

Modelin tek bir öznitelik ile eğitimi yapıldığında dahi yaklaşık aynı başarım sroku ile tahminler yapabildiği gözlemlenmiştir. Burda ortaya başarım skoru parametresinin model başarısını tam olarak yansıtamadığı durumu ortaya çıkmıştır. Problemin sebebi modelin aslında öğrenme işlemini doğru bir şekilde yapamıyor olmasıdır. Çünkü üzerinde çalışılan veri kümesi dengesiz veri kveri sayılarında dengesizlik olan bir veri kümesidir. Hedef öznitelikler (tahmini yapılmakta olan) arasındaki dengesizlik tabloda gösterilmiştir.

Veri Kümesinde Dengesizlik


Tablo 4 - Veri setinde bulunan kanser özniteliğinin değer dağılımı

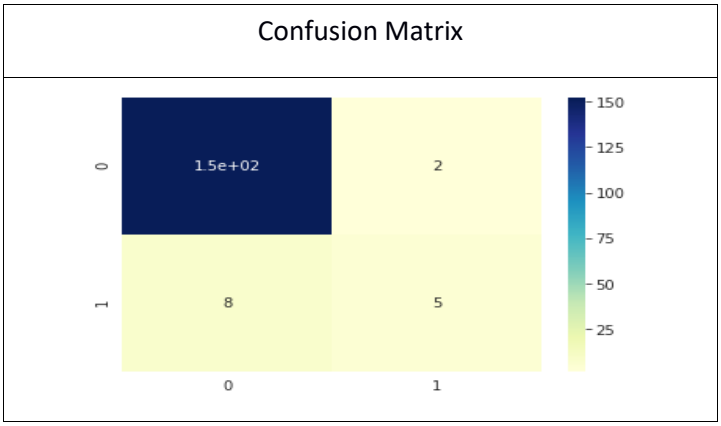
Bu kapsamda diğer metrik olan Recall (duyarlılık) metriği ile kanser olanları doğru tepsit etme oranın ne durumda olduğu incelenmiştir.

Recall (Duyarlılık)	0.38461538461538464

$$recall = \frac{true\ positives}{true\ positives\ + false\ negatives}$$

Tablo 5 - Recall (duyarlılık) değeri

Recall (duyarlılık) metriği sonucunda başarım skoru çok yüksek olan modelin aslında kanser olanları doğru tepsit etme oranının çok düşük olduğu saptanmıştır.



Tablo 6 - Tahmin ve gerçek sonuçlar arasındaki karmaşıklık matrisi

Karmaşıklık matrisinden de görüldüğü üzere test setinde bulunan 13 kanser hastasından sadece 5 tanesine model tarafından kanser tanısı konulmuştur. 8 tane kanser hastası kanser değil olarak nitelendirilmiştir. Bu durum aslında modelin çok yanlış karar verdiğinin göstergesidir.

Yapılan bu ilk deneyin amacı gerçek hayattaki verilerinin bir çoğunun dengesiz olmasından dolayı model güvenilirliği açısından başarım skoru parametresinin yetersizliğini göstermektir.

3.2 Veri Ön İşleme

3.2.1 Veri Gözlemleme

Rahim Ağzı Kanseri Veriseti Her Bir Öznitelik için Eksik Değer Sayıları	
Age	0
Number of sexual partners	26
First sexual intercourse	7
Num of pregnancies	56
Smokes	13
Smokes (years)	13
Smokes (packs/year)	13
Hormonal Contraceptives	108
Hormonal Contraceptives (years)	108
IUD	117
IUD (years)	117
STDs	105
STDs (number)	105
STDs:condylomatosis	105
STDs:cervical condylomatosis	105
STDs:vaginal condylomatosis	105
STDs:vulvo-perineal condylomatosis	105
STDs:syphilis	105

STDs:pelvic inflammatory disease	105
STDs:genital herpes	105
STDs:molluscum contagiosum	105
STDs:AIDS	105
STDs:HIV	105
STDs:Hepatitis B	105
STDs:HPV	105
STDs: Number of diagnosis	0
STDs: Time since first diagnosis	787
STDs: Time since last diagnosis	787
Dx:Cancer	0
Dx:CIN	0
Dx:HPV	0
Dx	0
Hinselmann	0
Schiller	0
Citology	0
Biopsy	0

Tablo 7 - Veri setindeki eksik değer (missing value) sayıları

Veri seti üzene yapılan gözlem doğrultusunda veri setinde eksik değerlerin olduğu gözlemlenmiştir. Tabloda veri setinde her bir sütunda bulunan eksik değer sayısı gösterilmektedir. STDs: Time since first diagnosis ve STDs: Time since last diagnosis özniteliklerinde bulunan eksik değer sayısı tüm örnek sayısının 858 olduğu düşünüldüğünde dikkat çekici şekilde fazladır.

3.2.2 Eksik Değerler Üzerine işlemler

Veri kümemizde yer alan "STDs: Time since first diagnosis" ve "STDs: Time since last diagnosis" özniteliklerinde çok fazla eksik değer bulunduğu için gerçekçi yaklaşımdan kopmamak adına verisetinden çıkarılır.
Ayrıca içerisinde eksik **değer bulunduran herbir örnek veri setinten çıkarılarak** eksik değer tablosu aşağıdaki duruma getirilmiştir.

Rahim Ağzı Kanseri Veriseti Her Bir Öznitelik için Eksik Değer Sayıları	
Age	0
Number of sexual partners	0
First sexual intercourse	0
Num of pregnancies	0
Smokes	0
Smokes (years)	0
Smokes (packs/year)	0
Hormonal Contraceptives	0
Hormonal Contraceptives (years)	0

IUD	0
IUD (years)	0
STDs	0
STDs (number)	0
STDs:condylomatosis	0
STDs:cervical condylomatosis	0
STDs:vaginal condylomatosis	0
STDs:vulvo-perineal condylomatosis	0
STDs:syphilis	0
STDs:pelvic inflammatory disease	0
STDs:genital herpes	0
STDs:molluscum contagiosum	0
STDs:AIDS	0
STDs:HIV	0
STDs:Hepatitis B	0
STDs:HPV	0
STDs: Number of diagnosis	0
Dx:Cancer	0
Dx:CIN	0
Dx:HPV	0
Dx	0
Hinselmann	0
Schiller	0
Citology	0
Cancer	0

Tablo 8 - Yapılan işlem sonrası eksik değer sayıları

Yapılan **eksik değer çıkarma** işlemlerinin ardından veri setinde eksik değer bulundurmayan **34 özniteliğe** sahip **668 örnek** bulunmaktadır. Ayrıca literatür taraması sonucu hedef öznitelik değerleri olarak Biopsy değeri alınmıştır. Çalışma netincesinde bu özniteliğin adı Cancer olarak değiştirilmiştir.

3.3 Dengesiz Veri Kümesi İçin SMOTETomek Yaklaşımı

Yeniden örnekleme yaparak, dengesiz veri kümelerini daha dengeli hale getirebiliriz. Bunu yapmak için ilk yöntem azınlık sınıfına ait verileri çeşitli yöntemlerle arttırarak eşit sayıda veriye sahip sınıflar elde etmektir (oversampling).
Diğer yöntem ise ağırlıklı sınıfa ait verileri veri kümesinden çıkararak dengeli bir veri kümesi elde etmektir(undersampling).

Bu çalışmada oversampling ve undersampling kombinasyonu olan SMOTETomek (synthetic minority oversampling technique+Tomek Link) yaklaşımı ile veri kümesi dengelenmiştir.

SMOTETomek tekniği; imbalanced_learn kütüphanesinde bulunan, interpolasyon yöntemiyle sentetik veriler üreteren SMOTE oversample yöntemini ve üst üste gelen verileri (overlapping samples) temizleyen Tomek Link yöntemini birlikte kullanan bir tekniktir. Bu

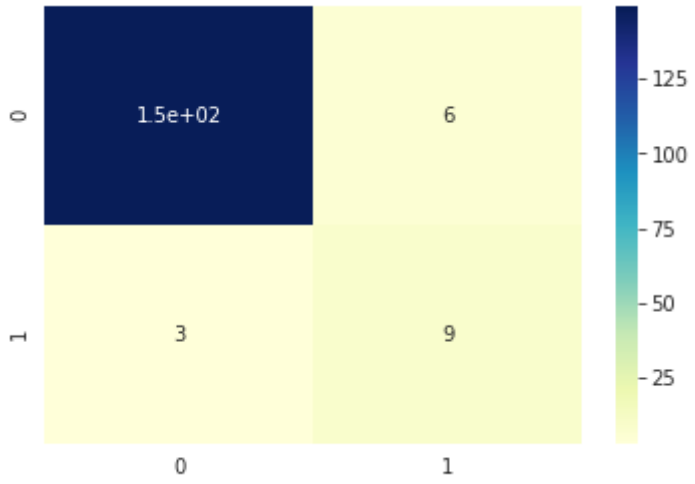
nedenle, SMOTETomek tekniği dengesiz veri seti problemini veri kümesinin yapısını bozmadan çözebilmektedir.

Train veri kümesine SMOTETomek yöntemi uygulanmadan önce ve uygulandıktan sonraki Kanserli hasta sayısı tabloda yer almaktadır.

	Kanserli	Kanserli Değil
SMOTETomek Uygulanmadan Önce	33	468
SMOTETomek Uygulandıktan Sonra	465	465

Tablo 9 - SMOTETomek metodu öncesi ve sonrası örnek sayısı

SMOTETomek yöntemi ile dengelenen veri kümesnin doğruluk skoru ve recall (duyarlılık) değerleri aşağıdaki tabloda yer almaktadır.



Accuracy score:	0.9461077844311377
Recall:	0.75

Tablo 10 - SMOTETomek metodu sonrası başarım skoru ve recall değeri

Yapılan dengeleme çalışması sonrası Recall (duyarlılık) metriğimiz bariz oranda artmıştır.

3.4 Özniteliklerin Önem Tespiti

Kullanılan rahim ağzı kanseri veri kümesi bir çok öznitelik buldundurmaktadır. Modelin tanımlanabilirliğini sağlama yaklaşımında, bu özniteliklerden hangisi modelin kanser tahmini yapması esnasında daha etkili hangisi daha az etkili bilinmesi modelin çalışma prenbisini anlamaya yardımcı olmaktadır.

Öznitelik önemi hesaplamanın bir çok yöntemi vardır. Bu çalışmada eli5 kütüphanesi tarafından sağlanan ‘Permutation Importance’ metodu kullanılmaktadır. Eli5 tarafından

sunulan Permutation Importance metodu, bir öznitelik mevcut olmadığında puanın nasıl azaldığını ölçerek kara kutu (black-box) problemleri için öznitelik önemlerini hesaplamayı sağlamaktadır.

Permutation Importance Metodu Sonucu

Weight	Feature
0.0647 ± 0.0418	Schiller
0.0072 ± 0.0048	STDs:syphilis
0.0060 ± 0.0000	STDs: Number of diagnosis
0.0048 ± 0.0048	STDs
0.0036 ± 0.0059	Num of pregnancies
0.0036 ± 0.0096	Age
0.0012 ± 0.0048	Hinselmann
0.0012 ± 0.0048	First sexual intercourse
0.0012 ± 0.0048	IUD
0.0000 ± 0.0076	Number of sexual partners
0 ± 0.0000	IUD (years)
0 ± 0.0000	Dx:HPV
0 ± 0.0000	STDs:condylomatosis
0 ± 0.0000	STDs:vaginal condylomatosis
0 ± 0.0000	Smokes (packs/year)
0 ± 0.0000	STDs:molluscum contagiosum
0 ± 0.0000	Dx:CIN
0 ± 0.0000	STDs:genital herpes
0 ± 0.0000	STDs (number)
0 ± 0.0000	STDs:cervical condylomatosis
… 13 more …	

Tablo 11 - Özniteliklerin Önem Sıralaması ve Ağırlıkları

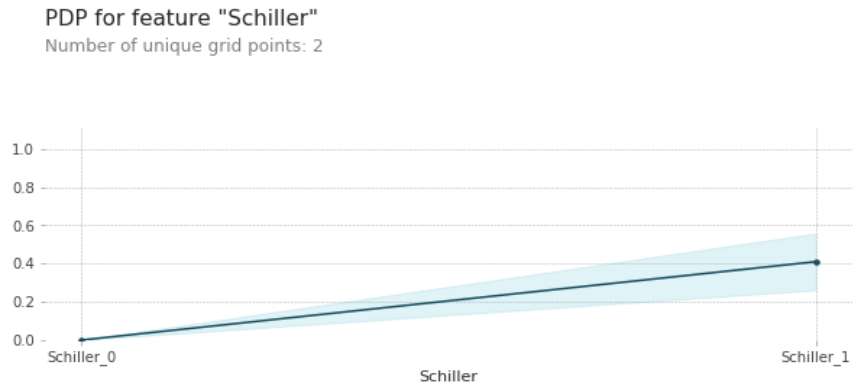
Bu çalışma ile modelin tahmin yaparken karar mekanizmasını özniteliklerin hangi derecede etkiledikleri analiz edilmektedir. Tabloda aşağıdan yukarı doğru gidildikçe karar mekanizmasına etki artmartadır. Örneğin model için en önemli öznitelik Schiller testi olduğu saptanmıştır.

3.5 Tek bir Özniteliğin Tahmine Etkisinin Nasıl Olduğunun Analizi

Bir önceki çalışmada modelin en önemli öznitelikleri saptanmıştır, ancak her bir özniteliğin karar makenizmasına olan etkisini incelemek kara kutu (black-box) problemine açıklık getiren yaklaşımlardan biridir. Örneğin 'Yaş' (Age) özniteliğinin karar mekanizması için önemli bir öznitelik olduğunu saptanmıştır fakat kanser şansının yaşla birlikte arttığını veya azaldığını bilmiyoruz. Bu örnekte olduğu gibi tek özniteliğin değerlerinin kendi içinde modelin tahminini nasıl etkilediğini bilmek model tanımlanabilirliğini arttırıcı faktörlerden biridir.

Bu çalışmada, tek bir özniteliğin model tahminini nasıl etkilediğini bilmek için “Kısmı Bağımlılık Grafikleri” adı verilen bir teknik kullanılmaktadır. Çalışma presibi ise tek bir tahmin yapmak yerine satırın bir değişkenini tekrar tekrar değiştirerek bir dizi tahmin yapma esasına dayanır. Örneğin rahim ağzı kanseri modeli için test verilerinden bir satır alır ve yaş gibi tek bir değişken değerini tekrar tekrar değiştirirek bir dizi tahmin yapar ve bu işlemi birden çok satır için yerine getirir. Bulunan tahmin değerlerinin sonuçlarının ortalamaları grafiklerde dikey eksende, özniteliğin değerleri yatak eksende tutularak üzerinde çalışılan özniteliteki değişim için model davranışı analiz edilir.

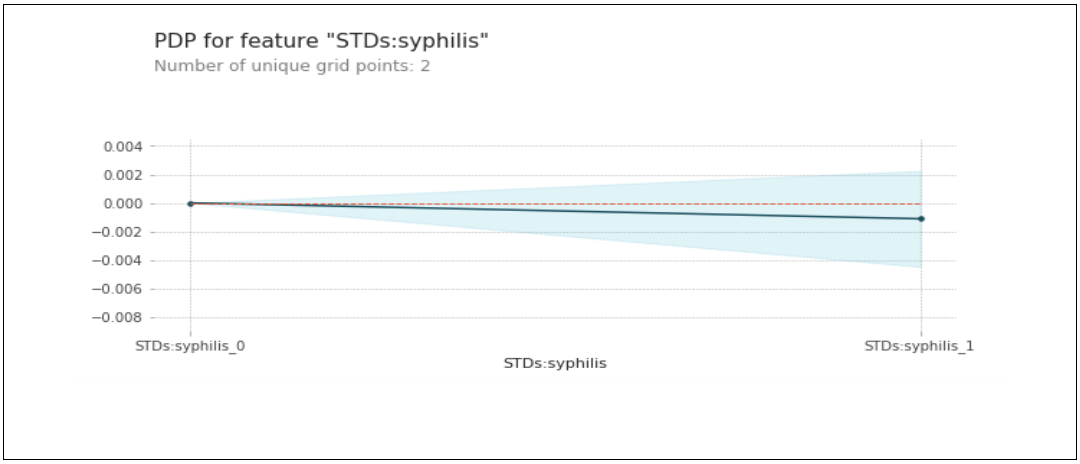
Çalışma için PDPBox kütüphanesi kullanılmıştır. Çalışma bulguları aşağıdaki tablolalarda yer almaktadır.

Schiller Testi


Tablo 12 - Schiller Testi Sonuçlarının Model Üzerinde Etkisi

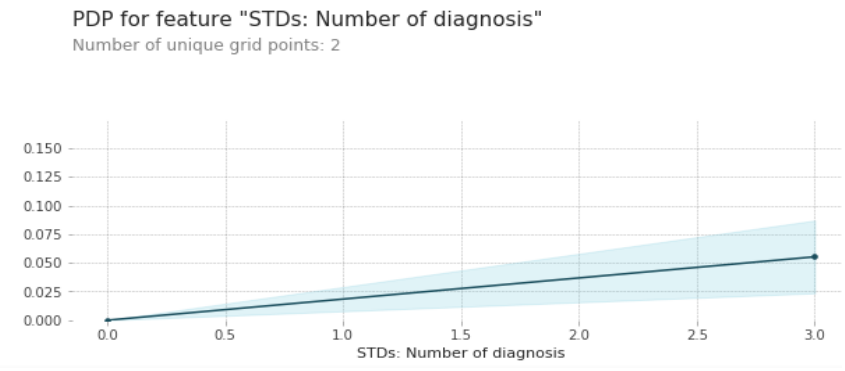
Bulgular Schiller testi sonucunun 1 (pozitif) olması durumunda kanser riskinin arttığını göstermektedir.

STD:syphilis



Tablo 13 - Syphilis Hastalığı bulunma durumunun modele etkisi

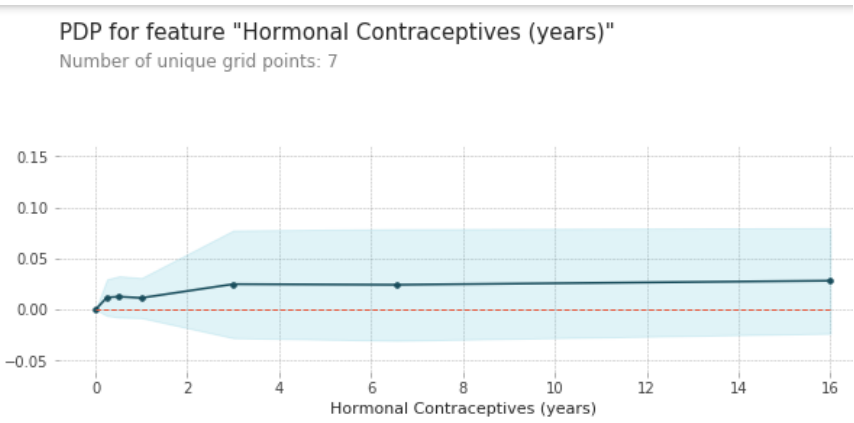
Bulgular Syphilis hastalığı olan bireylerde kanser riskinin data az olduğunu göstermektedir.

STDs: Number of diagnosis


Tablo 14 - Cinsel yolla bulaşan hastalık sayısının modele etkisi

Bulgular cinsel yolla bulaşan hastalık sayısı arttıkça kanser riskinin arttığını göstermektedir.

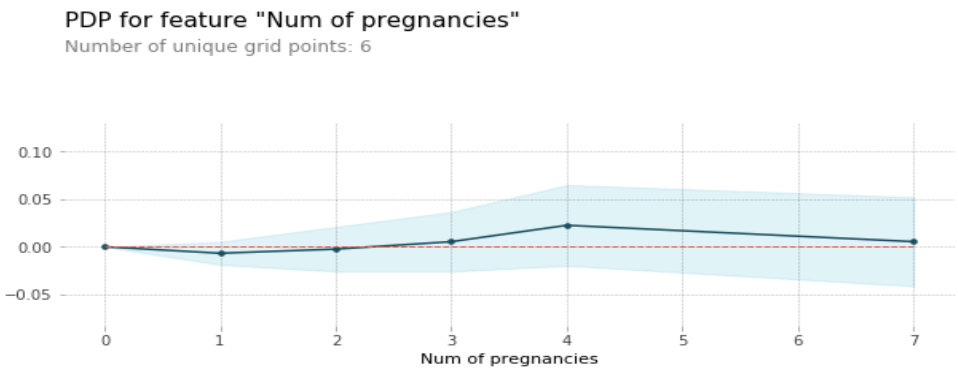
Hormonal Contraceptives (years)



Tablo 15 - Yıllık hormonal gebelik önleyici kullanımının modele etkisi

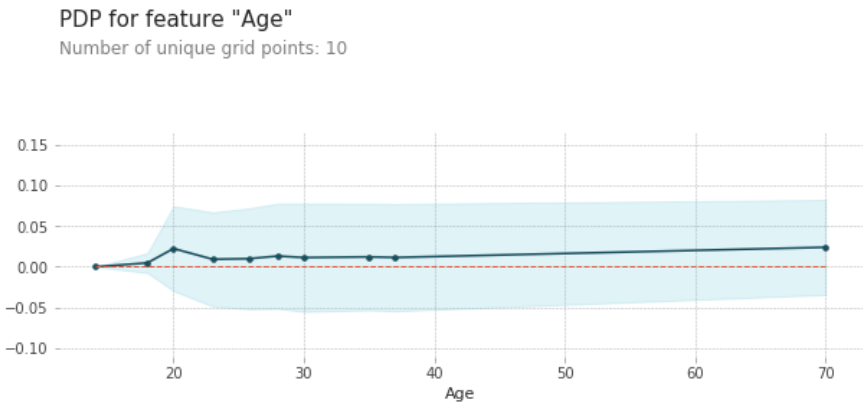
Bulgular yıllık kullanılan hormonal gebelik önleyici sayısının kanser riskini arttırdığını göstermektedir.

Number of Pregnancies



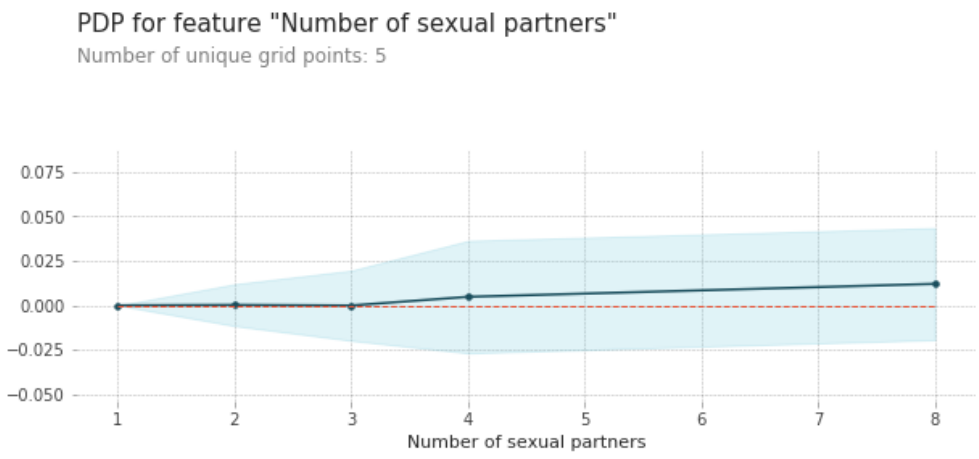
Tablo 16 - Hamilelik sayısının modele olan etkisi

Bulgular hamilelik sayısının 3'ten fazla olması durumunda kanser riskinin arttığını ve 4 kez hamilelik durumunun en riskli olduğunu göstermektedir.

Age


Tablo 17 - Yaştaki değişimin modele etkisi

Bulgular en riskli yaşın 20 olduğunu ve yaş arttıkça kanser riskinin attığını göstermektedir.

Number of Sexual Partners


Tablo 18 - Birlikte olunan kişi sayısının modele etkisi

Bulgular birlikte olunan kişi sayısı 3'ten fazla olduğu durumlarda kanser riskinin arttığını göstermektedir.

3.5 Yapılan Tahminin Analizi

Bu çalışmada modelin yaptığı bir tahmini hangi öznitelikleri ne kadar baz alarak yaptığının analizi yapılmaktadır. Bu analizler için SHAP değerlerinden faydalanılmaktadır. SHAP değerleri modelin yaptığı tahmini yorumlamak için kullanılan tekniklerden birisidir.

Örneğin; üzerinde çalışılan kanser tahmini modelinde, modelin belirli özniteliklere dayanarak birisinin kanser olduğunu tahmin ederse, her özniteliğin birlikte öngörüye nasıl katkıda bulunduğuna açıklık getirmetedir. Burada katkıda bulunmak, bir özniteliğin yapılmış kanser tahminini nasıl etkilediği anlamına gelmektedir.

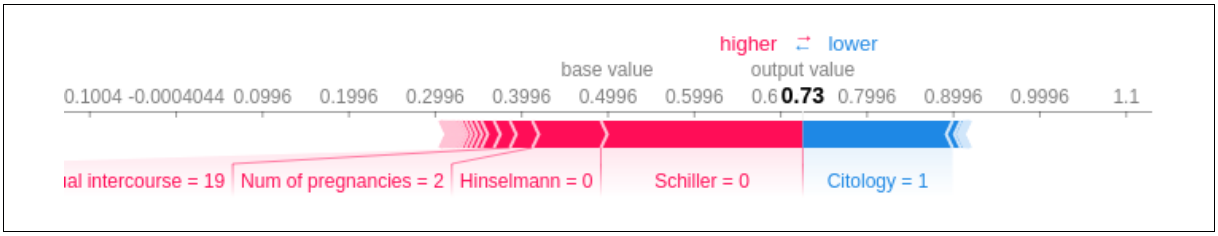
Çalışmada veri kümesinden rastgele seçilen iki örnek test olarak kullanılıp örneğin kanser olup olmadı tahmin edilmektedir daha sonra SHAP değerleri ile modelin neden bu şekilde tahminde bulunduğu analiz edilmektedir.

Ornek1

Ornek1 için tahmin Sonucu:	array([[0.726, 0.274]])

Tablo 19- Ornek1 için tahmin sonucu

Model, Ornek1'in %72.6 ihtimalle kanser olmadığını, %27.4 ihtimalle kanser olduğunu tahmin etmiştir. Yapılan tahminin neden bu şekilde sonuç verdiğini analiz etmek için SHAP değerleri aşağıda yer almaktadır.



Tablo 20 - Ornek1 tahmin sonucu için SHAP değerleri

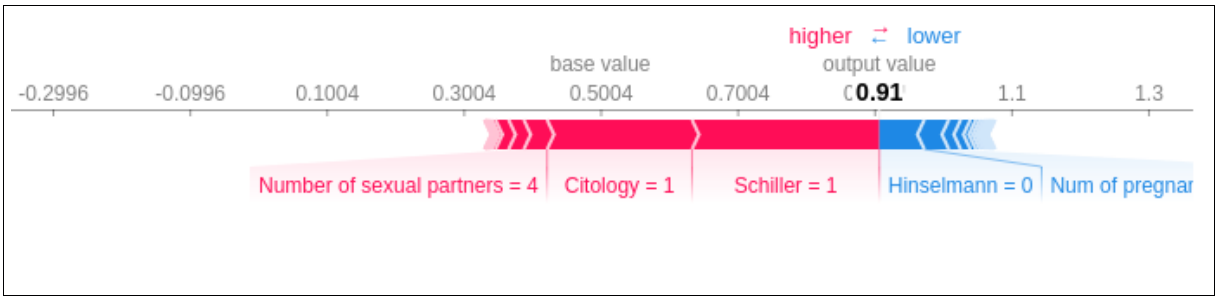
- Artan tahminlere neden olan öznitelik değerleri pembe renktedir, tahmini azaltan öznitelik değerleri mavi renktedir. Görsel boyutları özniteliğin etkisinin büyüklüğüni göstermektedir.
- Schiller = 0'ın modelin tahmin sonucu üzerinde en büyük etkiye sahip olduğu görülmektedir. Ayrıca Citology= 1 , sonucun azalması üzerinde çok büyük etkiye sahiptir. Bu durum, Schiller = 0 'ın Ornek1 in kanser olmama ihtimalini arttırdığı, Citology=1 'in ise kanser olmama ihitimalini düşürdüğü yani kanser ihtimalini arttırdığı olarak yorumlanmaktadır.

Ornek2

Ornek2 için tahmin Sonucu:	array([[0.092, 0.908]])

Tablo 21 - Ornek2 için tahmin sonucu

Model, Ornek2'nin %9.2 ihtimalle kanser olmadığını, %90.8 ihtimalle kanser olduğunu tahmin etmiştir.Yapılan tahminin neden bu şekilde sonuç verdiğini analiz etmek için SHAP değerleri aşağıda yer almaktadır.



Tablo 22 - Ornek2 tahmin sonucu SHAP değerleri

Bu analizde, Number of sexual partners = 4 , Schiller=1 ve Citology=1 modelin sonucunu arttırıcı şekilde etkilendiği gözlemlenmektedir. Bu durumun mantıklı olduğu daha önceki analizdeki tutarlılık ile anlaşılmaktadır. Ayrıca, Hinselmann = 0 ve Number of pregnancies = 1 modelin çıktısını yani kanser olma durumunun oranını azaltıcı etki yapmaktadır.

4)Sonuç

Bu çalışmada kara kutu probleminin çözümü ve model tanımlanabilirliğinin gerekliliği üzerinde durulmuş ve bu bağlamda deneyler yapılmıştır. Yapılan bu deneylerde modele olan güven kriterini sadece "Başarım Skoru" üzerine şekillendirmenin yanlış bir tutum olduğu, gerçek hayattaki dengesiz verilerle oluşturulmuş modellerde hataya sebep oluşturabilecek bir metrik olduğu ispatlanmıştır. Özellikle sağlık gibi hata kabulu olmayan bir sektör seçilip gerçek verilerle çalışılmıştır. Kritik alanlarda "Başarım Skoru" parametresinin yetersiz kaldığı ve modelin güvenilirliği konusunda kullanıcıyı yanıltabileceği ortaya konmuştur.

Rahim ağzı kanseri gibi kritik bir tahmin modeli için model tanımlanabilirliği ve anlaşılabilirliği adına değişik teknikler ve yaklaşımlar izlenerek modelin davranış analizi yapılmıştır. Yapılan analizlerle modelin karar mekanizmasına açıklık getirilip; hangi özniteliklerin ne şekilde bu karar mekanizmasına etki ettikleri gözlemlenmiştir.

Kaynakça:

- Classification of Cervical Cancer Dataset, Y. M. S. Al-Wesabi, Avishek Choudhury, Daehan Won Binghamton University, USA
- Data-Driven Diagnosis of Cervical Cancer With Support Vector Machine-Based Approaches, Hao Zhou
- Transfer Learning with Partial Observability Applied to Cervical Cancer Screening, Kelwin Fernandes, Jaime S. Cardoso, and Jessica Fernandes
- https://christophm.github.io/interpretable-ml-book/shap.html
- https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29
- https://www.kdnuggets.com/2018/12/machine-learning-explainability-interpretability-ai.html
- https://www.memorial.com.tr/saglik-rehberleri/rahim-agzi-kanseri-hakkindaki-gercekler-ve-tedavi-yontemleri/
- https://github.com/scikit-learn-contrib/imbalanced-learn/issues/589
- https://christophm.github.io/interpretable-ml-book/interpretability-importance.html
- https://christophm.github.io/interpretable-ml-book/shap.html
- https://pdpbox.readthedocs.io/en/latest/papers.html#notes-and-highlights

- https://eli5.readthedocs.io/en/latest/blackbox/permutation_importance.html
- https://www.researchgate.net/publication/335722569_SMOTETomek-based_Resampling_for_Personality_Recognition_July_2019
- https://imbalanced-learn.readthedocs.io/en/stable/auto_examples/under-sampling/plot_illustration_tomek_links.html
- https://medium.com/@g.canguven11/dengesi%CC%87z-veri%CC%87-k%C3%BCmeleri%CC%87-i%CC%87le-maki%CC%87ne-%C3%B6%C4%9Frenmesi%CC%87-63bbac5f6869
- https://github.com/SangeethaSA/Bank-churn-classification-SMOTETomek/blob/master/Bank%20churn%20data%20model%20-%20classification.ipynb
- https://towardsdatascience.com/understanding-a-black-box-896df6b82c6e