ALKOL DIŞI YAĞLIKARACİĞER HASTALIĞI

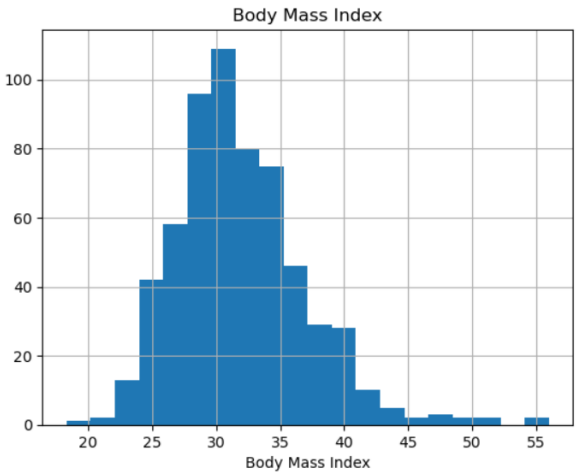
I.Giriş

Bu projede, tıbbi veri kümesi kullanılarak alkole bağlı olmayan yağlı karaciğer hastalığının sınıflandırılması amaçlanmıştır. Hangi hastaların bu hastalığa sahip olabileceğini öngörmek, tıbbi müdahaleleri erken aşamada planlamak ve hastalıkların erken teşhisini sağlamak amacıyla hastaların veri setindeki test değerleri kullanılarak bir makine öğrenimi modeli geliştirilmiştir. Biz projemizde hastaların değerleri sonucunda "Hastalık Türü" nün 1 hafif hastalık, 2 ağır hastalık olacak şekilde nasıl etkilendiğini inceledik.

Veri Setinin Temel Özellikleri:
Veri kümesi 605 hastadan alınan biyokimyasal ve fiziksel ölçümlerden oluşmaktadır. 605 satır ve 62 sütun bulunmaktadır. Veri seti integer ve float olmak üzere nümerik değişkenlerden oluşuyor. Ayrıca bazı eksik değerler bulunmakta. Hangi ölçümlerin kullanıldığı aşağıdaki tabloda verilmiştir:

İngilizce	Türkçe
Patient No.	Hasta No.
Age	Yaş
Gender (Female=1, Male=2)	Cinsiyet (Kadın=1, Erkek=2)
Height	Boy
Weight	Kilo
Body Mass Index	Vücut Kitle İndeksi
Waist Circumference	Bel Çevresi
Hip Circumference	Kalça Çevresi
Systolic Blood Pressure	Sistolik Kan Basıncı
Diastolic Blood Pressure	Diyastolik Kan Basıncı
Diyabetes Mellitus (No=0, Yes=1)	Diyabet Mellitus (Hayır=0, Evet=1)
Hypertension (No=0, Yes=1)	Hipertansiyon (Hayır=0, Evet=1)
Hyperlipidemia (No=0, Yes=1)	Hiperlipidemi (Hayır=0, Evet=1)
Metabolic syndrome (No=0, Yes=1)	Metabolik Sendrom (Hayır=0, Evet=1)
Smoking Status	Sigara Durumu
AST	AST
ALT	ALT
ALP	ALP
GGT	GGT
LDH	LDH
Total Bilirubin	Toplam Bilirubin
Direct Bilirubin	Direkt Bilirubin
Total Protein	Toplam Protein
Albumin	Albümin
Total Cholesterol	Toplam Kolesterol
Triglycerides	Trigliserit
HDL	HDL
LDL	LDL
Microalbumin Spot Urine	Mikroalbumin Spot İdrar
Microalbumin/Creatinine Ratio	Mikroalbumin/Kreatinin Oranı
TSH	TSH
CK	CK
Leukocyte	Lökosit
Hemoglobin	Hemoglobin
Trombosit	Trombosit
Mean Corpuscular Volume	Ortalama Eritrosit Hacmi
Mean Platelet Volume	Ortalama Trombosit Hacmi
PT	PT
INR	INR
Vitamin D	Vitamin D
Ferritin	Ferritin
Ceruloplasmin	Seruloplazmin
C Peptide	C-Peptid
Glucose	Glukoz
Insulin	İnsülin
HOMA	HOMA
Insulin resistance according to HOMA	HOMA'ya Göre İnsülin Direnci
Uric Acid	Ürik Asit
BUN	BUN
Creatinine	Kreatinin
Hemoglobin - A1C	Hemoglobin A1C
Steatosis	Steatoz
Activity	Aktivite
Fibrosis	Fibrozis
NAS score according to Kleiner	Kleiner'a Göre NAS Skoru
NAS score>=4 and Fibrosis>=2	NAS Skoru>=4 ve Fibrozis>=2
Fibrosis status	Fibrozis Durumu
Significant Fibrosis	Belirgin Fibrozis
Advanced Fibrosis	İleri Fibrozis
Cirrhosis	Siroz
Diagnosis according to SAF	SAF'ye Göre Tanı
Type of Disease	Hastalık Türü

Kodda değişkenlerin veri seti üzerindeki dağılımları şekildeki gibi görselleştirilmiştir ve bunların ortalama, standart sapma gibi istatistiksel özellikleri açıklanmıştır.



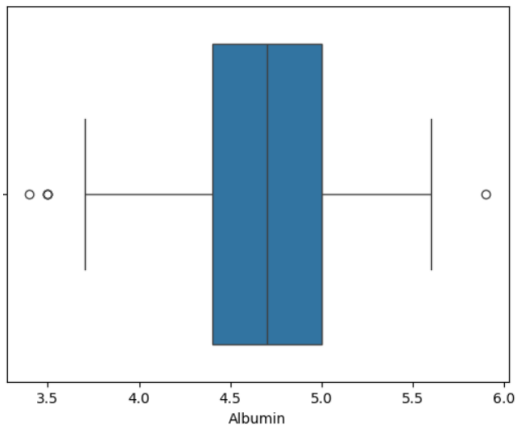
Type of Disease (Mild illness=1, Severe illness=2)	Microalbumin/Creatinine Ratio
1	59.790
2	94.237

Değerler ile hafif veya ağır hastalığa yakalanma arasındaki ilişki araştırılmıştır. Örneğin yukarıdaki çıktıda görüldüğü üzere hastalığı hafif geçirenlerde Mikroalbumin/Keratinin oranı 59.7 iken ağır hastalık geçirenlerde bu oran 94.2.

II.GELİŞME

VERİ ÖNİŞLEME VE SINIFLANDIRMA

Öncelikle eksik verileri ve bunların veri setindeki oranını tespit ettik. Eksik değer sayısı 300'den fazla olan sütunları kaldırdık. Daha sonra kalan eksik değerleri sütunun dağılımını bozmamak için her sütunun modunu kullanarak doldurduk. Bu sayede veri setinde artık eksik değer kalmadı.



Daha sonra sütun bazında alt ve üst sınırları hesaplayarak aykırı değerleri tespit ettik ve bunları kutu grafikleriyle görselleştirdik. Varsayılan olarak %10 ve %90'lık çeyreklik değerlerini kullandık. Aykırı değerlerin yerine sınır değerlerini koyduk. Yani alt sınırdan küçük olan değerleri alt sınıra, üst sınırdan büyük olan değerleri üst sınıra eşitledik.

Standartlaştırma: Verisetinde bulunan özelliklerin (özellikle sayısal olanların) aynı ölçekte olmasını sağlamak için normalizasyon işlemi gerçekleştirildi. Çünkü bazı makine öğrenimi algoritmaları, farklı ölçeklere sahip verilerle çalışırken zorluk yaşar. Örneğin, bir sütun 0-1 arasında, diğer sütun 0-1000 arasında değer alıyorsa, büyük değerlere sahip sütun modele daha fazla ağırlık verebilir.

Eğitim ve Test Verisi: Daha sonra modelin performansını değerlendirmek için veri setini %70 eğitim ve %30 test seti olmak üzere ikiye ayırdık. Eğitim seti, modeli eğitmek için kullanılır. Test seti, eğitilen modeli bağımsız bir veriyle değerlendirmek için kullanılır.

Makine Öğrenmesi Alogritmaları: K-Nearest Neighbors (KNN), Bayes, Decision Tree, Random Forest, Gradient Boosting (XGBoost) makine öğrenimi modellerini kullanarak problemdeki model performanslarını Accuracy, Precision, Recall, F1 Score, ROC-AUC sonuçlarıyla değerlendirdik.

	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Decision Tree	1.000	1.000	1.000	1.000	1.000
Random Forest	1.000	1.000	1.000	1.000	1.000
Gradient Boosting (XGBoost)	1.000	1.000	1.000	1.000	1.000
Support Vector Machine (SVM)	0.989	1.000	0.988	0.994	0.998
K-Nearest Neighbors (KNN)	0.984	0.994	0.988	0.991	0.998
Naive Bayes	0.978	1.000	0.976	0.988	1.000

Sonuçlar kullanılarak algoritmaların karşılaştırılması:

Accuracy (Doğruluk): Tüm veri üzerinden yapılan doğru tahminlerin oranını ifade eder. Karar Ağacı, Random Forest ve XGBoost modelleri %100 doğruluk (accuracy) elde etmiştir. Bu, bu modellerin test verisinde hiç hata yapmadığını gösterir.

Precision (Kesinlik): Pozitif olarak tahmin edilen sınıfların ne kadarının doğru olduğunu ölçer. Tüm modellerde precision %99.4 ile %100 arasında değişmektedir. Bu değer, yanlış pozitif (False Positive) sayısının oldukça düşük olduğunu ifade eder.

Recall (Duyarlılık): Gerçek pozitiflerin ne kadarının doğru tahmin edildiğini ölçer. Recall da %98.8 ve %100 arasında değişiyor, bu da modellerin pozitif sınıfları iyi bir şekilde tespit ettiğini gösterir.

F1 Score: Precision ve Recall'un harmonik ortalamasıdır. Dengesiz veri setlerinde özellikle

önemlidir. Burada Karar Ağacı, Random Forest ve XGBoost modelleri %100 F1 Score ile diğerlerinden daha iyi performans göstermektedir.

ROC-AUC: Modelin sınıflandırma performansını olasılık dağılımına dayalı olarak ölçer. Karar Ağacı, Random Forest ve XGBoost modelleri bu metrikte de mükemmel sonuç (%100) elde etmiştir.

Modeller çok yüksek doğruluk oranlarına ulaştı. Bu, modellerin veri seti üzerinde mükemmel performans gösterdiğini işaret ediyor. Bu durumun sıra dışı olduğunu düşünüp sorunları kontrol etmek için cross validation uyguladım. Sonuçlar aşağıdaki görselde:

	Mean F1 Score	F1 Score Std Dev
Naive Bayes	1.000	0.000
Decision Tree	1.000	0.000
Random Forest	0.997	0.005
Gradient Boosting (XGBoost)	0.997	0.005
Support Vector Machine (SVM)	0.995	0.005
K-Nearest Neighbors (KNN)	0.987	0.006

Naive Bayes ve Decision Tree:

Mean F1 Score: 1.000, yani mükemmel bir performans sergilemişler.

F1 Score Std Dev: 0.000, yani bu modeller her bir katmanda tamamen tutarlı sonuçlar üretmiş. Bu durum, verisetiyle uyumlarının çok yüksek olduğunu gösterir. Ancak, bu kadar yüksek performans aşırı öğrenmeye (overfitting) işaret edebilir.

Random Forest ve Gradient Boosting (XGBoost):

Mean F1 Score: 0.997, yani çok yüksek bir performans.

F1 Score Std Dev: 0.005, sonuçlar arasında düşük bir değişkenlik var, bu da modellerin tutarlı olduğunu gösterir.

Random Forest ve XGBoost, genellikle güçlü genel modellerdir. Burada da veriyle oldukça uyumlu çalıştıkları görülüyor.

Support Vector Machine (SVM):

Mean F1 Score: 0.995, yüksek performans.

F1 Score Std Dev: 0.005, sonuçlar oldukça tutarlı.

SVM'nin performansı, Random Forest ve XGBoost'a yakın. Ancak belki biraz daha düşük esneklik sergileyebilir.

K-Nearest Neighbors (KNN):

Mean F1 Score: 0.987, diğer modellere göre bir miktar daha düşük.

F1 Score Std Dev: 0.006, diğer modellere kıyasla biraz daha fazla değişkenlik var. Bu, modelin bazı katmanlarda performansının daha az tutarlı olduğunu gösterebilir.

Genel Değerlendirme:

Naive Bayes ve Decision Tree mükemmel sonuçlar göstermiş olsa da, bu sonuçlar aşırı öğrenme riskini işaret ediyor olabilir.

Random Forest, Gradient Boosting (XGBoost), ve SVM modelleri hem yüksek performans göstermiş hem de tutarlı çalışmış, bu da genelleme kabiliyetlerinin yüksek olabileceğini gösterir.

KNN ise göreceli olarak daha düşük performans sergilemiş, bu da bu veri setine diğer modeller kadar iyi uyum sağlayamadığını gösterebilir.

Sonuç olarak: Eğer genelleme kabiliyeti önemliyse, Random Forest veya Gradient Boosting gibi ensembel yöntemler tercih edilebilir.

Random Forest ve XGBoost Algoritmaları

Veri setinde 62 değişken bulunmakta. Bunlar arasında biyokimyasal ölçümler, demografik bilgiler, tanılar ve klinik sonuçlar yer alıyor. Random Forest ve XGBoost, yüksek boyutlu veri setlerinde etkili şekilde çalışabilir. Bu algoritmalar, özellikler arasında önemli olanları seçerek modelin karmaşıklığını ve hesaplama maliyetini azaltır. Özellikle, karar ağaçları temelli yapılar, önemli değişkenleri otomatik olarak belirleme yeteneğine sahiptir.

Veri setindeki değişkenler arasında güçlü etkileşimler ve doğrusal olmayan ilişkiler olabilir (örneğin, Bel Çevresi, Vücut Kitle İndeksi, Glikoz gibi değişkenlerin birlikte çalışarak Hastalık Şiddetini belirlemesi). Random Forest ve XGBoost, doğrusal olmayan ilişkileri iyi öğrenebilir ve değişkenler arasındaki etkileşimleri modele entegre edebilir.

Örneğin Type of Disease (Hafif Hastalık=1, Şiddetli Hastalık=2) sınıfındaki dağılım dengesiz olabilir. XGBoost ve Random Forest, ağırlıklı öğrenme yöntemlerini kullanarak azınlık sınıfını dikkate alabilir ve bu sınıfta da yüksek performans gösterebilir.

Her iki algoritma da Accuracy, Precision, Recall, F1 Score ve ROC-AUC gibi metriklerde maksimum performans sergilemiştir. Özellikle Precision ve Recall gibi metriklerde yüksek puanlar, yanlış pozitif ve yanlış negatif oranlarının düşük olduğunu gösterir. Bu, kritik hastalık sınıflandırmalarında hayati öneme sahiptir. ROC-AUC skorunun 1.000

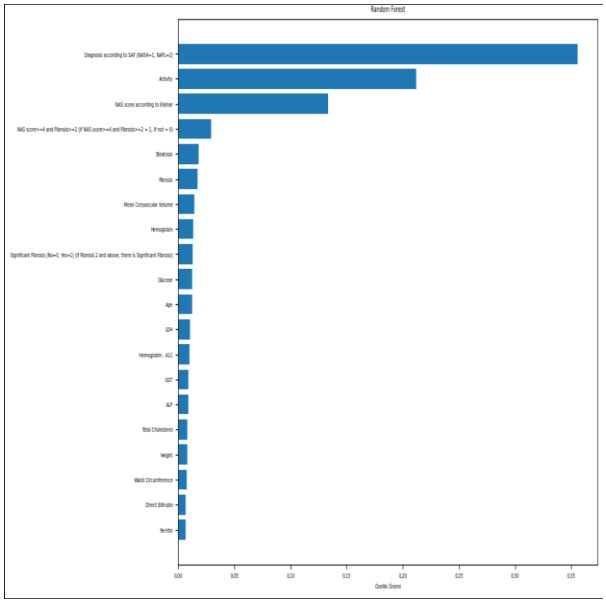
olması, modellerin sınıflar arasında mükemmel ayrım gücüne sahip olduğunu gösterir.

Random Forest, değişken önemini kolayca çıkarabilir, bu da doktorların veya sağlık çalışanlarının hangi özelliklerin daha önemli olduğunu anlamalarına yardımcı olur (örneğin, Bel Çevresi, LDL, Glikoz, Fibrozis Seviyesi gibi değişkenler hastalık şiddetini etkileyebilir).

XGBoost, SHAP değerleri aracılığıyla her bir özelliğin bireysel bir tahmin üzerindeki etkisini açıklayabilir. Örneğin, bir hasta için modelin tahmin ettiği "şiddetli hastalık" kararına hangi değişkenlerin ne kadar katkıda bulunduğunu görmek mümkündür.

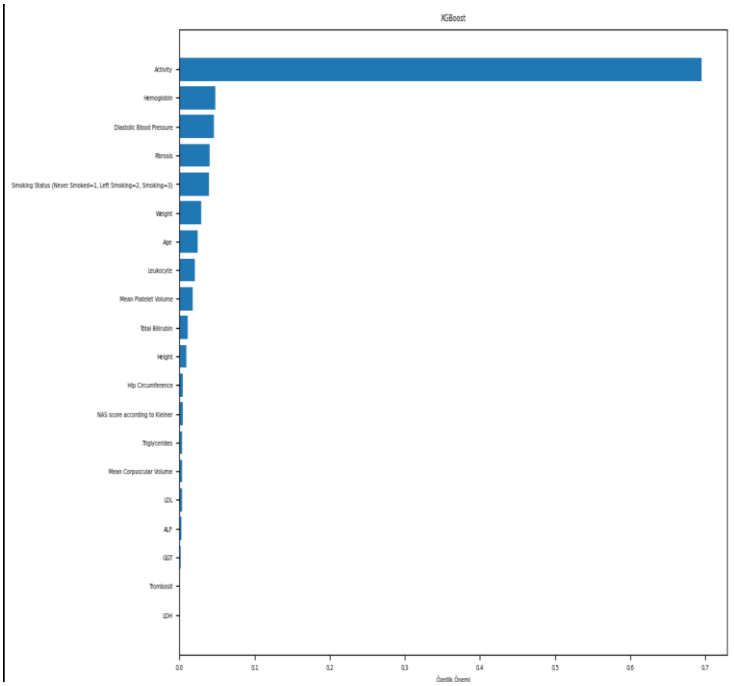
Makine öğrenimi modelinin hangi özelliklere daha fazla ağırlık verdiğini (yani hangi özelliklerin tahmin üzerinde daha etkili olduğunu) analiz etmek için özellik önemlerini görselleştirdik.

Random Forest Özellik Önemi:



Yukarıdaki tablo random forest modelinin hangi özelliklere ne kadar önem verdiğini gösteriyor. En önemli özellikler sırasıyla: "Diagnosis according to SAF" (NASH veya NAFL ayrımı), "Activity" "NAS score according to Kleiner". Diğer özellikler, öneme göre azalarak sıralanmış: Ferritin, Direkt Bilirubin, ve Bel Çevresi. Özellik önemleri açıkça görüldüğü için, doktorlar teşhislerde en önemli özelliklere daha fazla odaklanabilir.

XGBoost Özellik Önemi:



Activity özelliği, model için en yüksek öneme sahiptir. Bu, modelin tahmin yaparken en fazla bu özelliğe dayandığını gösteriyor. Yani aktivite seviyesinin hedef değişkenle güçlü bir ilişki içinde olduğu söylenebilir. Kişinin aktivite seviyesi ile hastalığı ağır geçirme üzerinde belirleyici bir faktör var. Özellikler arasındaki önem farkı oldukça belirgin. Örneğin, Activity'nin önem değeri diğerlerinden çok daha yüksek, bu da bu özelliğin veri setindeki hedef değişkeni (y) açıklamada kilit bir rol oynadığını gösteriyor.

SONUÇ

```
Sonuç Karşılaştırma:
```

	Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
0	Random Forest	1.000	1.000	1.000	1.000	1.000
1	XGBoost	1.000	1.000	1.000	1.000	1.000

Her iki model de çok yüksek doğruluk oranlarına ulaştı. Yani modeller veri seti üzerinde kusursuz çalışıyor. Ancak bunun aşırı öğrenme ya da başka herhangi bir problem teşkil etmediğini cross validation uygulayarak doğruladık. Sonuçların doğruluğu sayesinde özellik önemlerini kullanarak hangi test değerlerinin hastalığın "hafif" ya da "ağır" olduğunu belirlemede kritik rol oynadığı ortaya çıkabilir. "Hafif" hastalık durumundaki hastalar erken teşhis edilerek yaşam tarzı değişiklikleri veya tedavilerle "ağır" hastalık durumuna geçişleri önlenebilir. "Ağır" hastalık durumundaki hastalar için algoritmalar, bu grupta öne çıkan özellikleri inceleyerek daha yoğun tedavi gerektiren hasta gruplarını belirleyebilir.

Kaynakça

https://www.youtube.com/playlist?list=PLK8LlaNiWQOuTQisICOV6kAL4uoerdFs7

https://www.veribilimiokulu.com/

https://www.purestorage.com/knowledge/what-is-data-preprocessing.html

https://www.kaggle.com/hakankocakk/code/

https://github.com/RegaipKURT/Machine-Learning-Python