İKTİSADİ VE İDARİ BİLİMLER FAKÜLTESİ YÖNETİM BİLİŞİM SİSTEMLERİ BÖLÜMÜ

İLAY DOĞA AYHAN 33046094986

VERİ MADENCİLİĞİ

Dr. Öğr. Üyesi Nur Kuban TORUN

MEME KANSERİ KARAR AĞACI VE NEVEİ BAYES SINIFLANDIRMASI

Bilecik Şeyh Edebali Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, Yönetim Bilişim Sistemleri Bölümü

ÖZET: Kanserin temel sebeplerinden birisi de strestir. Stres oluşan hücreleri farklı bir boyuta çevirir. Hücrelerin yüksekliğini, çeper kalınlıklarına etki eder. Kanser günümüzün hastalığı olarak bilinmektedir. Büyük veriyi inceleyerek bunu analiz edebiliriz. Sigara, alkol, fazla hamur işi tüketimi, yaşam tarzı, obezite, radyasyon, katkı maddeleri en temel etkenlerdendir. Yazımızda bunların açıklaması olacaktır. Karar ağacı ve Naive Bayes yöntemleriyle kanser hücresinin nasıl büyüdüğünün analizini yaptık. Karar ağacı; hedefe ulaşma olasılığı en yüksek olan stratejiyi belirlemeye yardımcı olmak için kullanılan bir yöntemdir. Naive Bayes ise; sınıflandırması olasılık ilkelerine göre tanımlanmış bir dizi hesaplama ile sisteme sunulan verilerin sınıfını yani kategorisini tespit etmeyi amaçlar.

ANAHTAR KELİMELER: Karar ağacı, Naive Bayes, memekanseri

1.GİRİŞ

Meme kanseri, kadınları etkileyen en yaygın kanserdir. Her yaşta ortaya çıkabilir, ancak 60 yaş üstü kadınlarda daha sık görülür. Meme taraması (mamogram olarak da bilinir), meme kanserini erken teşhis etmenin en iyi yollarından biridir. Meme kanseri erken bulunursa, başarılı bir şekilde tedavi edilmesi ve hayatta kalma şansınızı artırması daha olasıdır.40'lı yaşlarınızda veya üzerindeyseniz, taramanın risklerini ve faydalarını doktorunuzla görüşün. Herhangi bir yaşta, herhangi bir meme semptomu veya memenizin görünümünde ve hissinde bir değişiklik fark ederseniz, gecikmeden doktorunuzu görmeniz önemlidir. Meme kanseri kadınlarda görülen kanserlerin %33'ünü oluşturmaktadır. Tüm kanser hastalarının ise %20'sini tehdit etmektedir. Günümüzde her 8 kadından 1'i hayatı boyunca meme kanseriyle karşı karşıya kalma riskiyle yaşamaktadır. Meme dokusu içinde süt kanalları içerisinde oluşan kanser hücreleridir. Meme kanserlerinin yüzde 80'i invaziv duktal karsinomdur. Invaziv duktal karsinom, meme kanserinin süt kanallarında ortaya çıktığını gösterir. Meme kanserinin yüzde 20'si de invaziv lobüler karsinomdur. Bu türde ise meme kanseri süt kanallarında değil, süt bezlerinde gelişir. Meme kanserine neden olan hücrelerin çoğalması ve büyümesi oldukça zaman alır. Ancak çoğaldıktan sonra hücreler lenf ve kan yoluyla vücudun diğer organlarına yayılabilir. Meme kanserinde en önemlisi kanserin kan ve lenf yolu ile diğer organlara yayılmadan tanının konmasıdır. Bu aşamada konulan bir tanı ile tedavi oranı çok yüksektir. Bu nedenle meme kanserinde erken teşhis çok önemlidir. Meme kanserini tetikleyen unsurlar arasında fazla stres ve unlu ürün tüketimidir. Günlük hayatta tüketimlerimize dikkat etmemiz gerekmektedir. Hücre boyutu; anormal hücreler yüksek riskin işaretçisidir. Bu, LCIS olan kadının, ilerde her iki memesinde de yayılabilen kanser olma riskinin yüksek olduğu anlamına gelir. Tümör boyutu 2 cm ya da daha küçüktür ve kanser hücreleri meme dışına (lenf bezlerine) sıçramamıştır.

Hücre yüksekliği; Kütle oluşumuna yol açan hücre birikimi, tümör olarak tanımlanır. Meme kanserinde bu durum sıklıkla meme dokusunda yer alan süt kanallarında veya süt bezlerinde yer alan hücrelerin hızla çoğalmasına bağlı olarak gelişir. Bu bölgelerde çoğalmaya başlayan kanser hücreleri, tümör denen kütleyi oluşturur. Çeper kalınlığı; meme muayenesinde belli

olur. Doku boyutu; teknik görüntülerle ortaya çıkar. Hücre boyutu; çok önemli bir teşhistir. Meme kanserinin evrelerini belirler. Meme Kanserinin evreleri; 4 evreden oluşmaktadır.

Evre-1; Tümör boyutunun 2 cm'den daha küçük olmasıdır ve tümör başka bir yere yayılmamıştır.

Evre-2; Tümör boyutunun 2 ve 5 cm. arasında olmasıdır. Fakat burada bazı alt gruplarında tümör hücreleri koltuk altı bezlerine miktarı fazla olmamakla birlikte yayılım yapabilir.

Evre-3; Tümörün 5 cm'den büyük olmasıdır ya da 5 cm'den küçük tümör olup koltuk altı bezlerinde fazla sayıda paket halinde tutulum olması, göğüs kasları duvarına tutulum olmuş olması, meme başını tutmuş olması ya da boyundaki lenf nodlarına doğru yayılım yapmış olması evre-3 olarak tanımlanır.

Evre-4 ise meme kanserinin başka organlara sıçramış olmasıdır.

Meme Kanserinden Korunmanın Yolları; Kadınların ilk olarak kendilerine bakmaları gerekmektedir. Sigara, alkol vb zararlı maddelerden uzak durmaları gerekiyor. Türk toplumumuzun geleneği olan hamur işini kenara bırakmaları ve sağlıklı ürünler yemeleri gerekmektedir . Hamur işleri kanseri tetikleyen temel maddelerin başında gelmektedir. Meme kanserini besleyip büyütmek istiyorsanız unlu ürünler tercih edebilirsiniz. Düzenli kiloyu korumaya dikkat etmelisiniz. Fazla kilo her zaman zararlıdır. Kadınlar genelde kilolarında dikkat etseler bile zayıflığın da sağlık olduğunu düşünürler. Fakat insanın kendi kilosu demek zayıf olmak demek değildir. Akdeniz usulü beslenmek meme kanserinden korunmanıza yol açabilir. Yeşilliklerle beslenerek bir tık daha korunma yoluna gidebilirsiniz .Bir diğer korunma yolu da düzenli spor yapmaktır. Düzenli spor yapmak hem omurga sağlığınızı hem de kilonuzun normalleşmesini sağlamaktadır. Normal kilo kanserden korunmanın temel başlığıdır. Sağlıklı beslenmek güçlü adımlar demektir.

Kadınsal ilaçlardan uzak durmak ise ayrı bir başlıktır. Farklı hormonsal ürünler kullanmak, içindeki etki maddeleri meme kanserine kapı açar. Hormonları değiştirmek, onların yapısıyla oynamak asla ve asla sağlıklı olmayan olayların başında gelir. İleri yaştaki kadınlar, menopoz sonrasında başlanan hormon yerine koyma tedavileriyle (HRT) östrojen ve progesteron gibi kadınlık hormonları alıyorlar. Östrojenin yanı sıra progesteron da içeren kombine tedavilerin, sadece östrojen içeren tedavilerden daha riskli olabileceği düşünülüyor. Bu nedenle tıbbi gerekçeler olmadan hormon kullanılmaması gerekiyor.

Meme Kanseri İlaçla Önlenebilir Mi? Meme kanseri açısından yüksek risk grubunda değerlendirilen kadınlarda önleyici hormon tedavisine başlanabiliyor. Bu amaçla tamoksifen ve raloksifen etken maddelerini içeren ilaçlar kullanılıyor. Çalışmalar, yüksek risk grubunda olan kadınlardaki meme kanseri görülme sıklığının, bu tür ilaçlarla yüzde 50'ye varan oranda azaldığını gösteriyor. Kanın daha kolay pıhtılaşması başta olmak üzere bazı yan etkileri olan bu ilaçlara, klinik kontrol sonrasında hekim tarafından başlanıyor.

Yukarıda açıklamış olduğumuz Meme Kanseri algoritmasını "KARAR AĞACI VE NAIVE BAYES SINIFLANDIRMASI" ile açıklayacağız. Bu yazımızda bizlere; R programlama, veri madenciliği, karar ağacı algoritması, NAVIE BAYES sınıflandırması yardımcı olmuştur.

1.1VERİ MADENCİLİĞİ NEDİR?

Büyük veriden anlamlı ve yararlı bilgilerin elde edilmesi sürecidir. Verinin büyüyen bir hızı vardır. Bu hıza ulaşmak veriyi yakalamak önceden uzun ve kara düzen bir süreçti. Şu anda ise Big data yani büyük veri analizi veri madenciliği sayesinde kolay bir hal almıştır. Büyük veri sayesinde dev bir bilgi birikimi parmaklarımızın ucunda. Volüm, Velocıyt, Variety; hacim, hız, çeşitlilik. Bunlar büyüyen verinin ana başlıklarıdır. Çeşitlilik farklı kaynaklardan farklı şekilde gelebilir. Büyük veride mesaj, ses, görüntü, metin veya veri tabanı dosyaları bulunabilir. Big data da kontrol paneli bulunur; metinler, postalar, hangi sayfalar, kaç dakikada ne yaptığımız, sosyal medya paylaşımlarımız, nereye ne zaman gittiğimiz? Bunların hepsini anlık olarak yapabilirler. Fakat bunlar gelecek değil şu an yaşadığımız alanda yer almakta.

Veri madenciliği ise bu büyük veriyi yöneten anlamlandıran kısımdır. Kimin hangi bilgiye ihtiyacı varsa oradan onu almasına yardımcı olur. Veri madenciliği bir süreçtir. Veri yığınları arasında soyut kazılar yaparak veriyi ortaya çıkarmanın yanı sıra, bilgi keşfi sürecinde örüntüleri ayrıştırarak süzmek ve bir sonraki adıma hazır hale getirmek de bu sürecin bir parçasıdır. Veri madenciliğinin süreçleri de vardır;

1.Problemin tanımlanması

2.Veriyi anlama

3.Veri Hazırlama

4. Model Kurma

5.Model Performans Değerlendirme Ölçütleri

6. Yayılım

AÇIKLAYALIM

1.Problemin tanımlanması; ilk olarak problemimizi bilmemiz gerekiyor. Problemi bilmeden tanımadan analizini gerçekleştiremeyiz. Problemi iyi şekilde bulmaya çalışmalıyız. Denenmiş olan problemleri tekrar piyasaya sürmemeliyiz.

2.Veriyi anlama; Nominal: kategorik tipteki değişkenlerdir. Birbirine göre üstünlük belirtmez. Örneğin; meslek grupları. Akademisyen. Binary(ikili): evet hayır şekliden örnek verilebilir. Nominal(kategorik) niteliktedir. Farkı 2 değer almasıdır. Örneğin: evet- hayır. Kadın-erkek. Ordinal(sıralı): kategorik nitelikler sıralı yazılıyorsa ya da birbirilerine üstünlük taslıyorsa sıralı olarak adlandırılır. Örneğin: düşük-orta-yüksek, işletmenin küçük-orta-büyük olması. İnteger(tamsayı): tamsayı tipinde aritmetik işlemler yapabildiğimiz nitelikler. Örneğin : kardeş sayısı.Interval-scaled(aralık ölçeği): Selsiyus-Fahrenayt. Ratio-scaled(Oran ölçeği): Sıfır gerçek yokluğu ifade eder. Örneğin, hastanın kilosu, boyu. Veriyi anlamlayız , verinin derdini çözmek için onu ilk önce anlamamız gerekmektedir. Psikolojide ; empati gibi.

3.Veri Hazırlama; Toplama, değer biçme, birleştirme ve temizleme örneklem seçimi, dönüştürme. Bunları veri hazırlanmasında bize yardımcı olan temel başlıklardır. Toplama; ilk olarak bizim kafamızda taslak olarak çıkmasını sağlayan birikimdir. Değer biçme; önemlilik arz eden unsurlara bakar ve araştırma yaparız. Birleştirme; elimizde toplanan verileri birleştirip uzaktan bakarız. Neredeyiz ve ne kadarız? Son olarak yavaş yavaş dönüştürmeye başlarız.

Veriyi hazırlamak uzaktan basit bir işlem gibi gözükse de ince ayrıntılar burada meydana gelmektedir. Uç noktalar tespit edilir. Normallik dağılımına bakılır. Kutu grafiği ile birlikte aykırı noktalar bulunabilir. Tekrar eden gözlemler veri setinden çıkarılmalıdır. Sebebi ise veri setini fazla şişirmesidir. Kayıp değerlerin tanımlanması. Nasıl tespit edilir? Eğer nitelik nümerik ise, tüm niteliğe ait ortalama kayıp değer yerine yazılır. Eğer değer kategorik ise, en çok tekrar eden değerin yerine yazılması gerekir. Eğer kayıp değer sayısı çok fazla ise ilgili değişken analizden tamamıyla çıkarılabilir.

Normalizasyon; Bir veri setinde ele alınan değişkenlerin, değişim aralıkları birbirinden oldukça uzak olduğunda, analizin olumsuz etkilenmemesi için yapılan, veri değerlerini küçük bir aralığa indirgeme işlemine normalizasyon denir.

Min – Max Normalizasyon Yöntemi; Veriyi 0 ile 1 arasındaki sayısal değerlere dönüştürmek için uygulanır. Bu yöntem veri içindeki en büyük ve en küçük sayısal değerin belirlenerek diğerlerini buna uygun biçimde dönüştürme esasına dayanmaktadır.

z-Score Normalizasyon Yöntemi; Z-skoru temel bir standart skor türüdür (aritmetik ortalama = 0.0 ve standart sapma 1.0). Standart skor, istenilen değerin ve popülasyon aritmetik ortalamasının farkının, popülasyon standart sapmasına bölünmesiyle elde edilir. Bu dönüştürme işlemine, standartlaştırma veya normalleştirme denir.

4. Model Kurma; Modelin bulunması için çok fazla deneme yanılma yöntemi gerekmektedir. Doğru test verisini bulmak oldukça zordur. Deneme yanılma yöntemi süreklilik gösterir. Sürekli olarak denemek ve doğruluk oranına bakmamız gerekmektedir. Denetimli ve denetimsiz öğrenme olarak bakabiliriz.

5.Model Performans Değerlendirme Ölçütleri; Makine Öğrenimi algoritmaları, belirli bir durum için kullanılır ve benzer durumdaki bilinmeyenler için sonuç tahmin etmek için kullanın. Bu algoritmalar arasında problemi çözmek için kullanın. Bu problemlerden ikisi regresyon ve sınıflandırma problemleridir.

Regresyon, temel olarak değişkenler arasındaki ilişkiyi bulmak için istatistiksel bir yaklaşımdır. Makine öğreniminde, veri kümesinden elde edilen değişkenler arasındaki ilişkiye dayalı olarak bir olayın sonucunu tahmin etmek için kullanılır.

Sınıflandırma, veri öğelerinin ait olduğu sınıfı belirtir ve çıktıda sonlu ve ayrık değerler olduğunda kullanılır. Bir giriş değişkeni için bir sınıf öngörür. Bir veya daha fazla girdi verildiğinde, bir sınıflandırma modeli bir veya daha fazla sonucun değerini tahmin etmeye çalışacaktır.

Veri setinde hedef nitelik olması durumunda, danışanlı öğrenmenin kullanıldığı bir sınıflandırma probleminde, bir modelin performansının değerlendirilebilmesi için en sık kullanılan yöntemlerden biri kontenjans tablosu/olabilirlik çizelgesi/hata matrisi(confusion matrix) dayanmaktadır.

		GERÇEK		
		POZİTİF	NEGATİF	TOPLAM
TAHMİN	POZİTİF	Doğru Pozitif(DP)	Yanlış Pozitif(YP)	tPoz
	NEGATİF	Yanlış Negatif(YN)	Doğru Negatif(DN)	tNeg
	TOPLAM	poz	neg	m

KONTENJANS TABLOSU

- DP: Doğru pozitif. Gerçekte meme kanseri hastası olan hastalardan modelin meme kanseri hastası olarak tahmin ettiği hastaların sayısıdır.
- YP: Yanlış pozitif. Gerçekte meme kanseri hastası olmayan hastalardan modelin meme kanseri hastasıdır biçiminde tahmin ettiği hastaların sayısıdır.
- YN: Yanlış negatif. Gerçekte meme kanseri hastası olan hastalardan modelin meme kanseri hastası değildir biçiminde tahmin ettiği hastaların sayısıdır.
- DN: Doğru negatif: Gerçekte meme kanseri hastası olmayan hastalardan modelin meme kanseri hastası değildir biçiminde tahmin ettiği hastaların sayısıdır.

6.Yayılım; veri madenciliğinin son aşamasıdır. Burada sona gelinmiştir.

2.Verilerin Elde Edilmesi

Verilerimi GitHub adlı platformdan buldum.699 tane veri ve 11 tane değişken vardır. Meme kanserini anlatan ve hücrelerle ilgili bilgi veren bir veri setidir. Verimin kaynağı internettir.

3.Metodoloji

3.1Karar Ağaçları Sınıflandırma

Karar ağacı, bir kurum veya kuruluş tarafından tercihlerin, risklerin, kazançların ve hedeflerin anlaşılmasına yardımcı olan bir teknik türüdür. Hem sınıflandırma hem de regresyonda kullanılır. Bir dizi sorular sorarız ve burada bunların cevaplarını alırız. Tahmin ve sınıflandırmada bizim işimizi kolaylaştırır. Sinir ağları gibi metodolojiler olsa da karar ağaçları kolay yorumlanabilir. Karar ağaçları; maliyet açısından düşüktür ve yorumlanması kolaydır. Diğer türlere göre güvenilirliği yüksektir. Verinin sınıflandırılması, öğrenme ve sınıflama olarak 2 gruba ayrılır. Öğrenme aşamasında önceden eğitimi alınmış yani bilinen eğitim verisi, model oluşturmak için sınıflama algoritması tarafından analizi yapılır. Sınıflama basamağında ise test verisi, sınıflama kurallarına veya karar ağacın doğruluk payını öğrenmek amacıyla kullanılır. Eğer kurallar doğru ise verilerin sınıflandırma amacıyla kullanılır. Sınıflandırma yapmak zor bir eylem olsa da karar ağacı bunu kolaylıkla yol aldırır. Tercihlerin, risklerin, kazançların, hedeflerin tanımlanmasında yardımcı olabilen ve birçok önemli yatırım alanlarında uygulanabilen, birbirini izleyen şansa bağlı olaylarla ilgili olarak çıkan çeşitli **karar** noktalarını incelemek için kullanılan bir tekniktir. Karar ağacının büyük desteği de budur.

Eğitim verisindeki hangi alanların nasıl kullanılacağı belirlenmelidir. Bu alanda en yaygın kullanılan Entropi ölçümüdür. Entropi; ölçüsü ne kadar fazla ise o alanda oluşan kararsızlıklar fazladır. Kararsızlık ve açık olmayan olaylar hoş karşılanmaz. Bundan kaynaklı olarak karar ağacında Entropi ne kadar az ise o alanlar kullanılır. Karar ağaçlarında kullanılan birçok algoritma mevcuttur. ID3, C4.5, C5.0, CART, CHAID ve QUEST bunlara örnek olarak gösterilebilir.

3.2 NAVIE BAYES SINIFLANDIRMASI

Naïve Bayes sınıflandırması olasılık ilkelerine göre tanımlanmış bir dizi hesaplama ile, sisteme sunulan verilerin sınıfını yani kategorisini tespit etmeyi amaçlar. Naïve Bayes sınıflandırmasında sisteme belirli bir oranda öğretilmiş veri sunulur. Öğretim için sunulan verilerin mutlaka bir sınıfı/kategorisi bulunmalıdır. Öğretilmiş veriler üzerinde yapılan olasılık işlemleri ile, sisteme sunulan yeni test verileri, daha önce elde edilmiş olasılık değerlerine göre, verilen test verisinin hangi kategoride olduğu tespit edilmeye çalışılır. Elbette öğretilmiş veri sayısı ne kadar çok ise, test verisinin gerçek kategorisini tespit etmek o kadar kesin olabilmektedir.

Naïve Bayes sınıflandırma yönteminin birçok kullanım alanı bulunabilir fakat, burada neyin sınıflandırıldığından çok nasıl sınıflandırıldığı önemli. Yani öğretilecek veriler binary veya text veriler olabilir, burada veri tipinden ve ne olduğundan ziyade, bu veriler arasında nasıl bir oransal ilişki kurduğumuz önem kazanıyor.

4.Kullanılan Modelin Ölçütleri

	POZİTİF GERÇEK DURUM	NEGATİF GERÇEK DURUM
POZİTİF TAHMİN	GERÇEK POZİTİF(TP)	YANLIŞ POZİTİF(FP)
NEGATİF TAHMİN	YANLIŞ NEGATİF (FN)	GERÇEK NEGATİF (TN)

Doğruluk oranı: *Sınıflandırma yapmak önemlidir. Sınıflandırma yapmadığımız zaman oranları bilmeyiz doğruluk oranı da sınıflandırılmış etken sayısının toplam etken sayısına oranıdır.*

$$Doğruluk\ Oranı = \frac{TP + TN}{TP + FP + FN + TN}$$

Hata oranı: Hatayı bilmek ve analiz etmek gerekmektedir. Hata oranı önemlidir. Hata oranı demek; yanlış sınıflandırılmış etken sayısının toplam etken sayısına oranıdır.

$$Hata\ Oranı = \frac{FP + FN}{TP + FP + FN + TN}$$

Kesinlik oranı: Kesinlik oranı ana ve temel yapı taşıdır. Kesinlik netlik her zaman önemlidir. Pozitif olarak tahmin edilen doğru etken sayısının, pozitif olarak tahminlenen tüm etken sayısına oranıdır.

$$\boldsymbol{Kesinlik\ Oranı = \frac{TP}{TP + FP}}$$

Duyarlılık oranı*:* Oranlar burada aktif rol oynar ve doğru sınıflandırma pozitif etkenlere yol gösterir. Doğru sınıflandırılmış pozitif etken sayısının; toplam pozitif etken sayısına oranıdır. Duyarlılık oranı küçük bir etki gibi gözükse de anlamı büyüktür.

$$Duyarlılık\ Oranı = \frac{TP}{TP + FN}$$

Özgüllük(belirlilik oranı):*Doğru negatifler ; ters orantı gibi gözükebilir. Fakat burada ki açıklama* doğru negatiflerin toplam yanlışlara oranı şeklinde açıklanır.

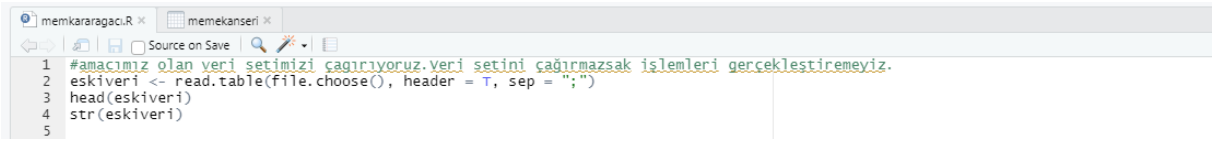
$$Özgüllük\ Oranı = \frac{TN}{TN + FP}$$

5.Uygulama

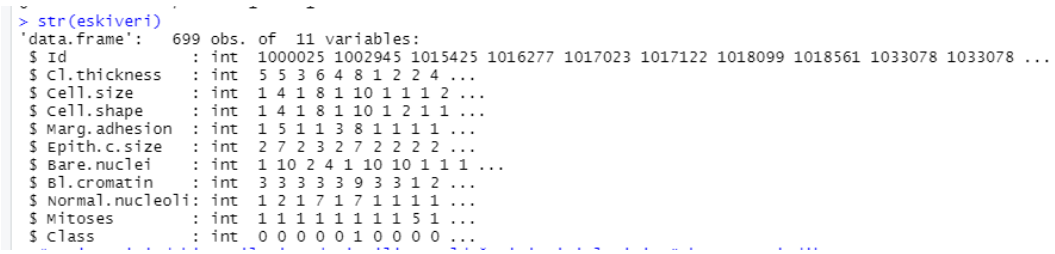
Veri setimizde meme kanserinde oluşan hücrelerin çeper kalınlığından, hücrenin boyutundan kanser olup olmadığını tespit ettik. Sınıflandırma algoritmalarından karar ağaçları ve naive bayes yöntemlerini kullandık. Bu işlemleri gerçekleştirirken R programlamadan yardım aldık ve kodlarımızı orada geliştirdik. Veri setimizde 699 gözlem değeri ve 11 adet değişken oluşmaktadır.699 tane gözlem değeri arasından 450 tane veriyi rastgele seçerek aldık. İnternetten bulduğumuz hazır veri setini csv formatıyla düzenleyerek R programına aktardık. Veri analizi için karar ağacı ve bayes ile uygun yöntemler getirilmiştir.

5.1 Karar Ağaçları Sınıflandırması

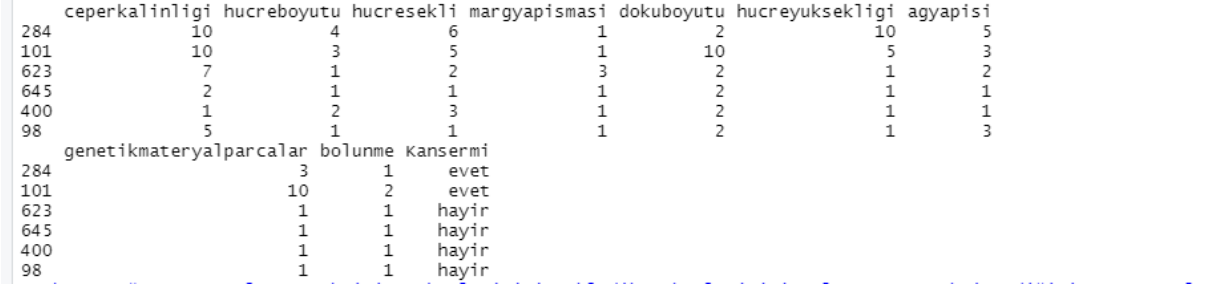
İlk olarak ; indirdiğimiz veri setimizi csv formatında excele aktarıyoruz .Sonra R programlamada veri setimizi çağırıyoruz. Bizim veri setimizin adı eski veri olarak kayıtlıdır. Eski veri yazdığımız zaman bizim verilerimizi çağırıyor. Head bizim verilerimizi görmemizi sağlar. Str ; Bir listenin sahip olduğu eleman sayısı, bu elemanların sınıfları ve değerleri “str()” fonksiyonu kullanılarak öğrenilir.



2.Daha sonrasında str(veri) komutuyla veri setimizin özetini görüntülüyoruz.



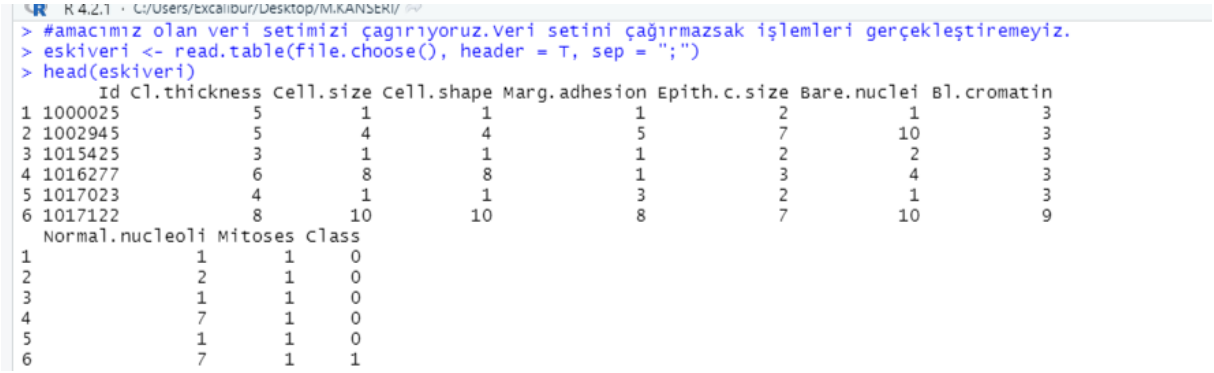
3. Bu kısımda ingilizce olan veri setimizin başlıklarını Türkçeye çevirip dil değişikliği yapıyoruz. Zorunlu bir eylem değildir. Fakat yapmamız bize anlamak için kolaylık sağlar.



	ceperkalinligi	hucreboyutu	hucresekli	margyapismasi	dokuboyutu	hucreyuksekligi	agyapisi
284		10	4	6	1	2	10	5
101		10	3	5	1	10	5	3
623		7	1	2	3	2	1	2
645		2	1	1	1	2	1	1
400		1	2	3	1	2	1	1
98		5	1	1	1	2	1	3

	genetikmateryalparcalar	bolunme	Kansermi
284	3	1	evet
101	10	2	evet
623	1	1	hayir
645	1	1	hayir
400	1	1	hayir
98	1	1	hayir

Burada verilerimiz İngilizce ;



	Id	Cl.thickness	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size	Bare.nuclei	Bl.cromatin
1	1000025	5	1	1	1	2	1	3
2	1002945	5	4	4	5	7	10	3
3	1015425	3	1	1	1	2	2	3
4	1016277	6	8	8	1	3	4	3
5	1017023	4	1	1	3	2	1	3
6	1017122	8	10	10	8	7	10	9

	Normal.nucleoli	Mitoses	Class
1	1	1	0
2	2	1	0
3	1	1	0
4	7	1	0
5	1	1	0
6	7	1	1

4.Bu kısımda meme kanseri olarak dosyamızın adını değiştiriyoruz. 699 tane olan verimizin 450 tanesini kullanmak için ind<-sample komutunu kullanıyoruz.

```
14 set.seed(1234)
15 ind <- sample(1:699,450)
16 memekanseri <- veri[ind,]
17 View(memekanseri)
18
```

```
> #veri setimizdeki verilerin adı ingilizce olduğu için isimlerini türkçeye çevirdik.
> veri <- eskiveri[, c("Id", "Cl.thickness","Cell.size","Cell.shape","Marg.adhesion","Epith.c.size", "Bare.nuclei", "Bl.cromatin", "Normal.nucleoli", "Mitoses", "Cl
> colnames(veri)<-c("Isım", "ceperkalinligi","hucreboyutu","hucresekli","margyapismasi","dokuboyutu","hucreyuksekligi", "agyapisi","genetikmateryalparcalar","bolunm
> veri <- veri[-1]
> set.seed(1234)
> ind <- sample(1:699,450)
> memekanseri <- veri[ind,]
> View(memekanseri)
> #burada factore dönüştürüyoruz
> memekanseri$kansermi<- as.factor(memekanseri$kansermi)
> levels(memekanseri$kansermi)<- c("0"="hayir","1"="evet")
> head(memekanseri)
```

5.Burada verilerimizi liste halinde görüyoruz.

```
> str(eskiveri)
'data.frame':   699 obs. of  11 variables:
 $ Id             : int  1000025 1002945 1015425 1016277 1017023 1017122 1018099 1018561 1033078 1033078 ...
 $ Cl.thickness   : int  5 5 3 6 4 8 1 2 2 4 ...
 $ Cell.size      : int  1 4 1 8 1 10 1 1 1 2 ...
 $ Cell.shape     : int  1 4 1 8 1 10 1 2 1 1 ...
 $ Marg.adhesion  : int  1 5 1 1 3 8 1 1 1 1 ...
 $ Epith.c.size   : int  2 7 2 3 2 7 2 2 2 2 ...
 $ Bare.nuclei    : int  1 10 2 4 1 10 10 1 1 1 ...
 $ Bl.cromatin    : int  3 3 3 3 3 9 3 3 1 2 ...
 $ Normal.nucleoli: int  1 2 1 7 1 7 1 1 1 1 ...
 $ Mitoses        : int  1 1 1 1 1 1 1 1 5 1 ...
 $ Class          : int  0 0 0 0 0 1 0 0 0 0 ...
```

6.Burada ise ; 0 ın hayır 1 in evet olduğunu yani meme kanseri mi değil mi ? Kararını veriyoruz.

```
levels(memekanseri$Kansermi)<- c("0"="hayir","1"="evet")


head(memekanseri)
```

7.Burada ise ; karar ağacımızı çalıştırmak için paketleri yükledik. Paketler bizim için her zaman önemlidir. Rweka , Rjava paketlerini kullandık. Karar ağacımızı çağırdık. Verilerimizle birlikte bize karar ağacımızı verdi. Karar ağacı için rWeka paketinin içerisindeki C4.5 algoritmasının J48() komutunu kullanıyoruz. Sonrasında da kurallarımızı görmek için print() fonksiyonunu kullanıyoruz.

```
27 #karar ağacımızı çalıştırmak için paketlerimizi yükledik.Paketlerimizi çalıştıramazsak istediğimiz sonucu alamayız.
28 library(Rweka)
29 kararagaci <- J48(Kansermi ~ ., data = memekanseri)
30 kararagaci
31 print(kararagaci)
32 summary(kararagaci)
33 plot(kararagaci)
34
```

8.Burada yukarıda yazdığımız kodun sonuçları gözükmektedir.

```
> #karar ağacımızı çalıştırmak için paketlerimizi yükledik.Paketlerimizi çalıştıramazsak istediğimiz sonucu alamayız.
> library(RWeka)
Warning message:
package 'RWeka' was built under R version 4.2.2
> kararagaci <- J48(Kansermi ~ ., data = memekanseri)
> kararagaci
J48 pruned tree
------------------

hucreboyutu <= 2
|   hucreyuksekligi <= 2: hayir (252.0/1.0)
|   hucreyuksekligi > 2
|   |   hucreyuksekligi <= 5
|   |   |   ceperkalinligi <= 6: hayir (18.0/1.0)
|   |   |   ceperkalinligi > 6: evet (2.0)
|   |   hucreyuksekligi > 5: evet (4.0)
hucreboyutu > 2
|   hucreyuksekligi <= 2
|   |   hucreboyutu <= 6
|   |   |   ceperkalinligi <= 7
|   |   |   |   dokuboyutu <= 4: hayir (15.0)
|   |   |   |   dokuboyutu > 4
|   |   |   |   |   hucreboyutu <= 5: evet (2.0)
|   |   |   |   |   hucreboyutu > 5: hayir (2.0)
|   |   |   ceperkalinligi > 7: evet (3.0)
|   |   hucreboyutu > 6: evet (9.0)
|   hucreyuksekligi > 2: evet (143.0/5.0)

Number of Leaves  :     10

Size of the tree :      19
```

Çıktımızda kurallarımız görülmektedir. Kurallar:

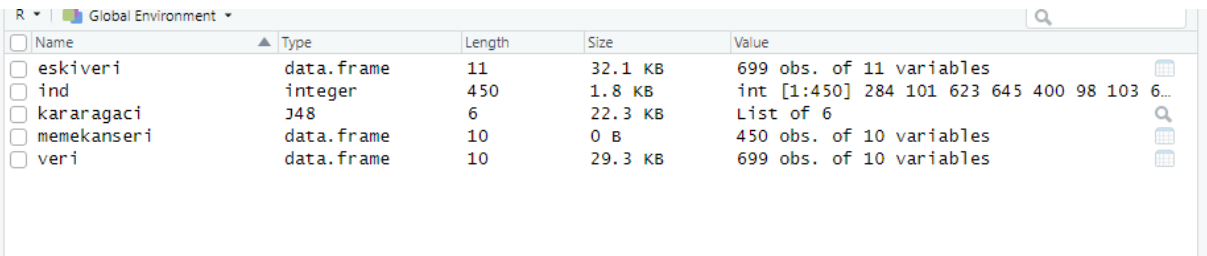
Eğer hücre boyutu 2 ye küçük eşit ise ve hücre yüksekliği 2 ye küçük eşit ise Hayır değildir.

Eğer hücre boyutu 2 den büyükse ve hücre yüksekliği 2 den büyükse ve Çeper kalınlığı 6 dan büyükse Evet kanserdir.

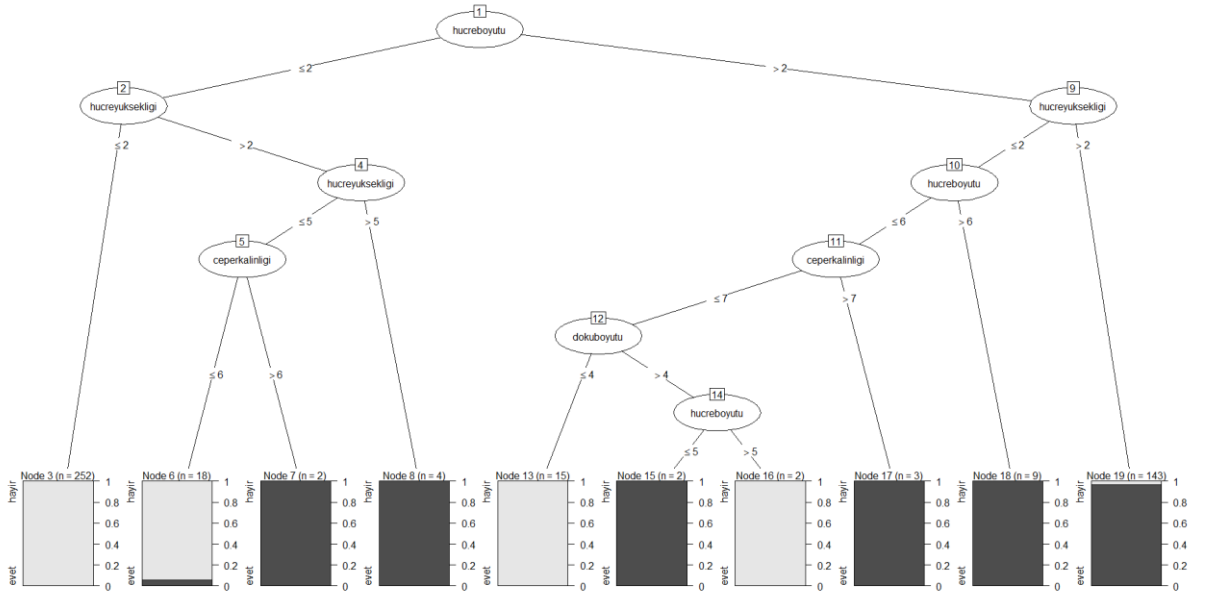
9.Veri setinin doğruluk oranını göstermekte ; Summary() komutuyla çağırdığımızda ise veri setimizin %98.4 oranında doğru olduğunu gösteriyor. Plot fonksiyonu ağacımızın grafiksel yapısını çizer.

```
> summary(kararagaci)

=== Summary ===

Correctly Classified Instances         443               98.4444 %
Kappa statistic                          0.9662
Mean absolute error                      0.0301
Root mean squared error                  0.1226
Relative absolute error                  6.5589 %
Root relative squared error             25.6154 %
Total Number of Instances              450

=== Confusion Matrix ===

   a   b   <-- classified as
 285   5 |   a = hayir
   2 158 |   b = evet
> plot(kararagaci)
```

10. Bu kısımda ise sırayla çıkan veri setlerimizin isimleri , değişen isimleri , içindeki bilgiler ,veri setimizin tablosu bu kısımda gözükmektedir.



11. SONUÇ ; Yukarıda kodlarını yazdığımız karar ağacının sonuçları , oranları gözükmektedir. Karar ağacı bizim hedefimizdir. Biz bu kadar kodu karar ağacının önümüze çıkması için uğraştık. Karar ağacı bizim programımızın şah damarıdır. Veri setimizin beynidir. Karar ağacını yorumlayıp olayı sonlandırıyoruz.

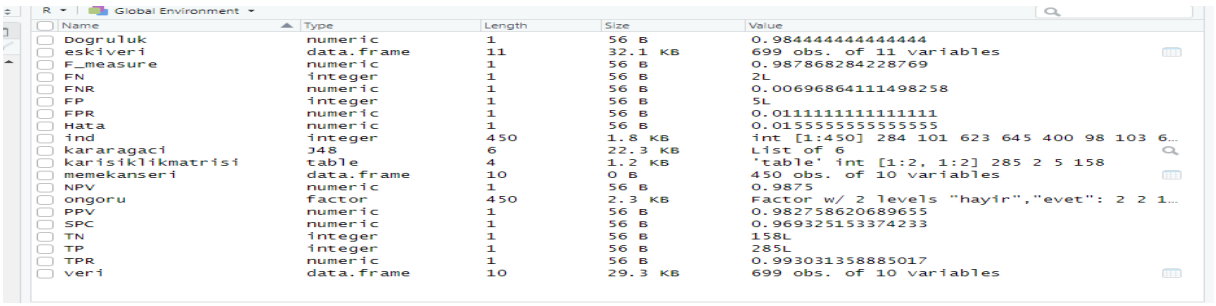


Modelimizin karışıklık matrisi görelim;

```
#ongoru yapalım

ongoru <- predict(kararagaci)
ongoru

#modelin karışıklık matrisi tabosuna bakalım
table(memekanseri$Kansermi, ongoru)
#isim koyalım
karisiklikmatrisi <- table(memekanseri[,10], ongoru)
karisiklikmatrisi

#modelin performans değerlendirme ölçütleri
#doğru pozitif
(TP <- karisiklikmatrisi [1])
#yanlış pozitif
(FP <- karisiklikmatrisi [3])
#yanlış negatif
(FN <- karisiklikmatrisi [2])
#doğru negatif
(TN <- karisiklikmatrisi [4])

#performans degerlendirme olcutleri
paste0("Dogruluk = ",(Dogruluk <- (TP+TN)/sum(karisiklikmatrisi)))
paste0("Hata = ",(Hata <- 1-Dogruluk))
#TPR=Duyarlilik orani
paste0("TPR= ", (TPR <- TP/(TP+FN)))
#SPC=Belirleyicilik orani
paste0("SPC= ", (SPC <- TN/(FP+TN)))
#PPV=kesinlik ya da pozitif ongoru degeri
paste0("PPV= ", (PPV <- TP/(TP+FP)))
#NPV=negatif ongoru degeri
paste0("NPV= ", (NPV <- TN/(TN+FN)))
#FPR=Yanlis pozitif orani
paste0("FPR = ", (FPR <- FP/sum(karisiklikmatrisi)))
#FNR=Yanlis negatif orani
paste0("FNR=",(FNR <- FN/(FN+TP)))
#F olcutu kesinlik ve duyarlilik olcutlerinin harmonik ortalamasi
paste0("F_measure = ", (F_measure <- (2*PPV*TPR)/(PPV+TPR)))
```

```
> #ongoru yapmamız gerekiyor
> ongoru <- predict(veri,memekanseri)
> table(memekanseri$Kansermi,ongoru)
   ongoru
      0   1
  0 279  11
  1   6 154
> #veri isimli veri setinin tum ongorulerine bakalım
> print(ongoru)
  [1] 1 1 0 0 0 0 0 0 0 0 0 1 1 0 1 0 0 1 0 0 0 1 1 0 0 0 0 1 0 0 1 1 0 0 0 1 0 1 0 1 0 1 1
 [56] 0 1 0 1 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 1 1 0 0 0 1 0 0 0 1 0 0 0 0 1 1 1 0 0 0 1 1 0 0 0 1 0 0 0 0 1 0 0 0
[111] 0 1 1 0 1 0 0 1 0 0 1 0 1 0 0 0 0 0 0 0 0 1 1 0 1 0 1 1 0 1 0 1 0 0 0 1 1 0 1 1 0 1 0 0 1 1 0 1 0 1 0 1 1 0 1
[166] 0 1 0 0 0 0 0 0 0 0 0 1 0 0 1 1 0 1 1 1 0 0 0 0 1 0 1 1 0 0 0 1 1 0 1 0 1 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0
[221] 0 0 0 1 0 1 1 1 1 0 0 0 0 1 0 1 1 1 0 1 0 0 0 0 0 0 0 1 1 1 0 1 0 1 0 0 1 0 0 0 1 0 1 0 0 1 0 1 0 0 0 0 1 0 0 0 0
[276] 0 1 0 1 1 0 1 0 1 0 0 0 1 0 0 0 1 1 0 1 0 1 0 0 0 0 1 1 0 0 1 0 0 0 0 0 0 1 1 1 0 1 0 1 0 1 1 0 1 0 0 1 0 1 0 0
[331] 1 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 1 1 0 0 0 0 0 0 0 0 0 1 0 1 0 1 1 1 0 0 0 1 0 1 0 1 0 0 1 0 0 1 0 0 0 0 1 0 1
[386] 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 1 0 0 0 1 0 1 1 0 1 0 1 1 1 1 1 0 0 0 1 1 0 0 1 1 0 0 0 0 0 0 0 0
[441] 0 1 1 0 1 0 0 1 1 0
Levels: 0 1
> #karisiklik matrisi
> karisiklikmatrisi <- table(memekanseri[,3],ongoru)
> karisiklikmatrisi
    ongoru
       0   1
  1  210   4
  2   38   4
  3   30  11
  4    6  26
  5    0  19
  6    1  18
  7    0  21
  8    0  17
  9    0   6
  10   0  39
> #dogruluk oranini bulmamız lazım onemli konu
> sum(diag(karisiklikmatrisi))/sum(karisiklikmatrisi)
[1] 0.4755556
```

```
> #doğru negatif
> (TN <- karisiklikmatrisi [4])
[1] 158
> #performans degerlendirme olcutleri
> paste0("Dogruluk = ",(Dogruluk <- (TP+TN)/sum(karisiklikmatrisi)))
[1] "Dogruluk = 0.984444444444444"
> paste0("Hata = ",(Hata <- 1-Dogruluk))
[1] "Hata = 0.0155555555555555"
> #TPR=Duyarlilik orani
> paste0("TPR= ", (TPR <- TP/(TP+FN)))
[1] "TPR= 0.993031358885017"
> #SPC=Belirleyicilik orani
> paste0("SPC= ", (SPC <- TN/(FP+TN)))
[1] "SPC= 0.969325153374233"
> #PPV=kesinlik ya da pozitif ongoru degeri
> paste0("PPV= ", (PPV <- TP/(TP+FP)))
[1] "PPV= 0.982758620689655"
> #NPV=negatif ongoru degeri
> paste0("NPV= ", (NPV <- TN/(TN+FN)))
[1] "NPV= 0.9875"
> #FPR=Yanlis pozitif orani
> paste0("FPR = ", (FPR <- FP/sum(karisiklikmatrisi)))
[1] "FPR = 0.0111111111111111"
> #FNR=Yanlis negatif orani
> paste0("FNR=",(FNR <- FN/(FN+TP)))
[1] "FNR=0.00696864111498258"
> #F olcutu kesinlik ve duyarlilik olcutlerinin harmonik ortalamasi
> paste0("F_measure = ", (F_measure <- (2*PPV*TPR)/(PPV+TPR)))
[1] "F_measure = 0.987868284228769"
>
```



Name	Type	Length	Size	Value
Dogruluk	numeric	1	56 B	0.984444444444444
eskiveri	data.frame	11	32.1 KB	699 obs. of 11 variables
F_measure	numeric	1	56 B	0.987868284228769
FN	integer	1	56 B	2L
FNR	numeric	1	56 B	0.00696864111498258
FP	integer	1	56 B	5L
FPR	numeric	1	56 B	0.0111111111111111
Hata	numeric	1	56 B	0.0155555555555555
ind	integer	450	1.8 KB	int [1:450] 284 101 623 645 400 98 103 6…
kararagaci	J48	6	22.3 KB	List of 6
karisiklikmatrisi	table	4	1.2 KB	'table' int [1:2, 1:2] 285 2 5 158
memekanseri	data.frame	10	0 B	450 obs. of 10 variables
NPV	numeric	1	56 B	0.9875
ongoru	factor	450	2.3 KB	Factor w/ 2 levels "hayir","evet": 2 2 1…
PPV	numeric	1	56 B	0.982758620689655
SPC	numeric	1	56 B	0.969325153374233
TN	integer	1	56 B	158L
TP	integer	1	56 B	285L
TPR	numeric	1	56 B	0.993031358885017
veri	data.frame	10	29.3 KB	699 obs. of 10 variables

5.2 Naïve Bayes Sınıflaması

1. csv formatında değiştirdiğimiz verileri R programına çağırıyoruz. Daha sonra Naive Bayesin kolu olan "Caret" ve "e1071" adlı paketlerimizi kütüphanemize yükleyip library() komutuyla çağırıyoruz.

```
#veri setimizden verilerimizi işlem yapabilmek için çağıralım.
eskiveri<- read.table(file.choose(), header = T, sep=";")

#naive bayes için e1071 ve caret paketlerimizi yüklememiz gerekiyor.Paketler önemli ayrıntılar
install.packages("e1071")
library(e1071)
library(caret)

```

2.Verilerimizi Türkçeye çeviriyoruz. Daha sonra veri setimizi parçalıyoruz.View komutu ile veri setimizi görüntüleyip liste haline getiriyoruz.

```

#verilerimizi türkçeye çeviriyoruz.
veri <- eskiveri[, c("Id", "Cl.thickness","Cell.size","Cell.shape","Marg.adhesion","Epith.c.size", "Bare.nuclei",
colnames(veri)<-c("Isım", "ceperkalinligi","hucreboyutu","hucresekli","margyapismasi","dokuboyutu","hucreyukseklig
#veri setini parçaladık
set.seed(1234)
ind <- sample(1:699,450)
memekanseri <- veri[ind,]
View(memekanseri)
```

Çıkan Sonuç

Filter

	Isım	ceperkalinligi	hucreboyutu	hucresekli	margyapismasi	dokuboyutu	hucreyukseklig	agyapisi	genetikmateryalparcalar	bolunme
1	1000025	5	1	1	1	2	1	3	1	
2	1002945	5	4	4	5	7	10	3	2	
3	1015425	3	1	1	1	2	2	3	1	
4	1016277	6	8	8	1	3	4	3	7	
5	1017023	4	1	1	3	2	1	3	1	
6	1017122	8	10	10	8	7	10	9	7	
7	1018099	1	1	1	1	2	10	3	1	
8	1018561	2	1	2	1	2	1	3	1	
9	1033078	2	1	1	1	2	1	1	1	
10	1033078	4	2	1	1	2	1	2	1	
11	1035283	1	1	1	1	1	1	3	1	
12	1036172	2	1	1	1	2	1	2	1	
13	1041801	5	3	3	3	2	3	4	4	
14	1043999	1	1	1	1	2	3	3	1	
15	1044572	8	7	5	10	7	9	5	5	
16	1047630	7	4	6	4	6	1	4	3	
17	1048672	4	1	1	1	2	1	2	1	
18	1049815	4	1	1	1	2	1	3	1	

Showing 1 to 18 of 699 entries, 11 total columns

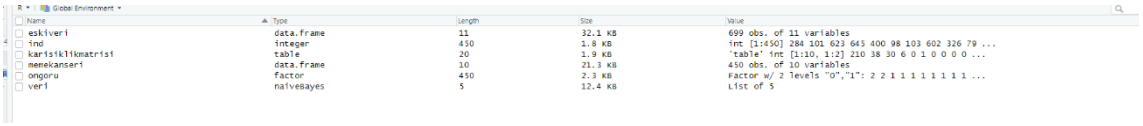
Tablomuzdan Elemanları çağırmak için str komutunu kullanıyoruz. Naive Bayes yöntemi için sayısal sütunları almamız gerekiyor bu yüzden 2:11 sütun aralığını seçiyoruz.

```
View(memekanseri)

#naive bayes için veri setimizin özet haline bakıyoruz.Emin olabilmek için.

View(memekanseri)
str(memekanseri)

#doğruluk oranı bulmamız için sayısal veriler almamız gerekiyor. bundan dolayı 2 ile 11 arasıkolonlarını getiriyor
memekanseri <-memekanseri[2:11]
str(memekanseri)
```

Modelimiz için e1071 paketinin içindeki n.b komutunu kullanıyoruz daha sonra öngörü yapmamız gerekiyor öngörü değierleri ile gerçek değerleri karşılaştırıyoruz. Veri isimli veri setinin tüm öngörülerine bakıyoruz daha sonra karşılık matrisine bakıp doğruluk oranını buluyoruz.

```
#MODEL için e1071 paketinin içindeki naivebayes komutunu kullanmamız gerekiyor.
veri <- naiveBayes(Kansermi~., memekanseri)
print(veri)

#ongoru yapmamız gerekiyor
ongoru <- predict(veri,memekanseri)

#ongoru değerleri ile gercek değerleri karsılastıralım

table(memekanseri$Kansermi,ongoru)

#veri isimli veri setinin tum ongorulerine bakalım
print(ongoru)

#karisiklik matrisi
karisiklikmatrisi <- table(memekanseri[,3],ongoru)
karisiklikmatrisi

#dogruluk oranini bulmamız lazım onemli konu
sum(diag(karisiklikmatrisi))/sum(karisiklikmatrisi)
```

Bu kısımda tablolarımızın içindekiler ve verilerin görüntü kısmı vardır.

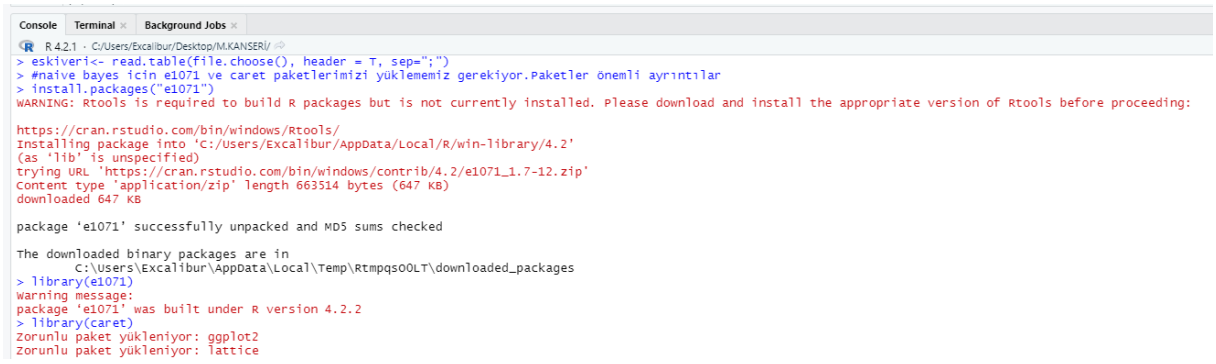


Name	Type	Length	Size	Value
eskiveri	data.frame	11	32.1 KB	699 obs. of 11 variables
ind	integer	450	1.8 KB	int [1:450] 284 101 623 645 400 98 103 602 326 79 ...
karisiklikmatrisi	table	20	1.9 KB	'table' int [1:10, 1:2] 210 38 30 6 0 1 0 0 0 0 ...
memekanseri	data.frame	10	21.3 KB	450 obs. of 10 variables
ongoru	factor	450	2.3 KB	Factor w/ 2 levels "0","1": 2 2 1 1 1 1 1 1 1 1 ...
veri	naivebayes	5	12.4 KB	List of 5

R programlamada n.b yaparken kullandığımız kodları sizlerle paylaştık. Kodların her birisi kendisine özgü özelliklere sahiptir. Özellikle paket kullanımlarına dikkat etmeliyiz. Önce ne istediğimizi bilmeliyiz ve sonra ne yapacağımıza karar vermeliyiz.

ÇIKTILAR

Navie Bayes için veri setimizi çağırdık ve çıktısını aldık. Paketlerimizi çalıştırmıştık ve çıktıları ;



```
Console  Terminal  Background Jobs
R 4.2.1 · C:/Users/Excalibur/Desktop/M.KANSERİ/
> eskiveri<- read.table(file.choose(), header = T, sep=";")
> #naive bayes icin e1071 ve caret paketlerimizi yüklememiz gerekiyor.Paketler önemli ayrıntılar
> install.packages("e1071")
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into 'C:/Users/Excalibur/AppData/Local/R/win-library/4.2'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/e1071_1.7-12.zip'
Content type 'application/zip' length 663514 bytes (647 KB)
downloaded 647 KB

package 'e1071' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Excalibur\AppData\Local\Temp\RtmpqsOOLT\downloaded_packages
> library(e1071)
Warning message:
package 'e1071' was built under R version 4.2.2
> library(caret)
Zorunlu paket yükleniyor: ggplot2
Zorunlu paket yükleniyor: lattice
```

Verilerimizi İngilizceden Türkçeye çevirmiştik ve verilerimizi str koduyla gördük ;

```
> #verilerimizi türkçeye çeviriyoruz.
> veri <- eskiveri[, c("Id", "Cl.thickness","Cell.size","Cell.shape","Marg.adhesion","Epith.c.size", "Bare.nuclei", "Bl.cromatin", "Normal.nucleoli", "Mitoses", "Class"
)]
> colnames(veri)<-c("Isım", "ceperkalinligi","hucreboyutu","hucresekli","margyapismasi","dokuboyutu","hucreyuksekligi", "agyapisi","genetikmateryalparcalar","bolunme", "K
ansermi")
> #veri setini parçaladık
> set.seed(1234)
> ind <- sample(1:699,450)
> memekanseri <- veri[ind,]
> view(memekanseri)
> view(memekanseri)
> str(memekanseri)
'data.frame':   450 obs. of  11 variables:
 $ Isım                : int  492268 1166654 1140597 1299596 1206314 1165790 1167471 1344449 743348 1133136 ...
 $ ceperkalinligi      : int  10 10 7 2 1 5 4 1 3 3 ...
 $ hucreboyutu         : int  4 3 1 1 2 1 1 1 2 1 ...
 $ hucresekli          : int  6 5 2 1 3 1 2 1 2 1 ...
 $ margyapismasi       : int  1 1 3 1 1 1 1 1 1 1 ...
 $ dokuboyutu          : int  2 10 2 2 2 2 2 1 2 2 ...
 $ hucreyuksekligi     : int  10 5 1 1 1 1 1 1 1 3 ...
 $ agyapisi            : int  5 3 2 1 1 3 3 2 2 3 ...
 $ genetikmateryalparcalar: int  3 10 1 1 1 1 1 1 3 1 ...
 $ bolunme             : int  1 2 1 1 1 1 1 1 1 1 ...
 $ Kansermi            : int  1 1 0 0 0 0 0 0 0 0 ...
```

Çalışan kodlarımız ;

```
> view(memekanseri)
> str(memekanseri)
'data.frame':   450 obs. of  11 variables:
 $ Isım                : int  492268 1166654 1140597 1299596 1206314 1165790 1167471 1344449 743348 1133136 ...
 $ ceperkalinligi      : int  10 10 7 2 1 5 4 1 3 3 ...
 $ hucreboyutu         : int  4 3 1 1 2 1 1 1 2 1 ...
 $ hucresekli          : int  6 5 2 1 3 1 2 1 2 1 ...
 $ margyapismasi       : int  1 1 3 1 1 1 1 1 1 1 ...
 $ dokuboyutu          : int  2 10 2 2 2 2 2 1 2 2 ...
 $ hucreyuksekligi     : int  10 5 1 1 1 1 1 1 1 3 ...
 $ agyapisi            : int  5 3 2 1 1 3 3 2 2 3 ...
 $ genetikmateryalparcalar: int  3 10 1 1 1 1 1 1 3 1 ...
 $ bolunme             : int  1 2 1 1 1 1 1 1 1 1 ...
 $ Kansermi            : int  1 1 0 0 0 0 0 0 0 0 ...
```

2 ile 11 arasındaki kolonları getirdik ;

```
> #doğruluk oranı bulmamız için sayısal veriler almamız gerekiyor. bundan dolayı 2 ile 11 arasıkolonlarını getiriyoruz ve tr fonksiyonu ile ekrana getirip kontrol ediyoru
z.
> memekanseri <-memekanseri[2:11]
> str(memekanseri)
'data.frame':   450 obs. of  10 variables:
 $ ceperkalinligi      : int  10 10 7 2 1 5 4 1 3 3 ...
 $ hucreboyutu         : int  4 3 1 1 2 1 1 1 2 1 ...
 $ hucresekli          : int  6 5 2 1 3 1 2 1 2 1 ...
 $ margyapismasi       : int  1 1 3 1 1 1 1 1 1 1 ...
 $ dokuboyutu          : int  2 10 2 2 2 2 2 1 2 2 ...
 $ hucreyuksekligi     : int  10 5 1 1 1 1 1 1 1 3 ...
 $ agyapisi            : int  5 3 2 1 1 3 3 2 2 3 ...
 $ genetikmateryalparcalar: int  3 10 1 1 1 1 1 1 3 1 ...
 $ bolunme             : int  1 2 1 1 1 1 1 1 1 1 ...
 $ Kansermi            : int  1 1 0 0 0 0 0 0 0 0 ...
```

Modelimiz için e1071 paketinin içindeki naivebayes komutunu kullandık ve çıktısı ;

```
> veri <- naiveBayes(Kansermi~., memekanseri)
> print(veri)

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
        0         1
0.6444444 0.3555556

Conditional probabilities:
   ceperkalinligi
Y       [,1]     [,2]
  0 2.97931 1.671950
  1 7.34375 2.428983

   hucreboyutu
Y       [,1]      [,2]
  0 1.303448 0.8671166
  1 6.525000 2.7241062

   hucresekli
Y       [,1]     [,2]
  0 1.482759 1.005889
  1 6.593750 2.555785
```

```
   margyapismasi
Y       [,1]      [,2]
  0 1.341379 0.9968869
  1 5.368750 3.2844629

   dokuboyutu
Y       [,1]      [,2]
  0 2.082759 0.8277493
  1 5.443750 2.5886712

   hucreyuksekligi
Y       [,1]     [,2]
  0 1.258621 1.018180
  1 7.475000 3.152218

   agyapisi
Y       [,1]      [,2]
  0 2.027586 0.9803949
  1 5.975000 2.2513448
```

```
   genetikmateryalparcalar
Y       [,1]     [,2]
  0 1.293103 1.119327
  1 5.706250 3.347010

   bolunme
Y       [,1]      [,2]
  0 1.058621 0.5580597
  1 2.781250 2.6556499
```

Öngörülerimizin sonuçları , karşılık matrisinin çıktısı ve doğruluk oranımız ;

```
> #ongoru yapmamız gerekiyor
> ongoru <- predict(veri,memekanseri)
> table(memekanseri$Kansermi,ongoru)
   ongoru
      0   1
  0 279  11
  1   6 154
> #veri isimli veri setinin tum ongorulerine bakalım
> print(ongoru)
  [1] 1 1 0 0 0 0 0 0 0 0 0 1 1 0 1 0 0 1 0 0 0 1 1 0 0 0 0 0 1 0 0 1 1 0 0 0 1 0 1 1 0 1 0 1 0 1 1
 [56] 0 1 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 1 1 0 0 0 1 0 0 0 1 0 0 0 0 1 1 1 0 0 0 1 1 0 0 0 1 0 0 0 0 0 1 0 0 0
[111] 0 1 1 0 1 0 0 1 0 0 1 0 1 0 0 0 0 0 0 0 0 0 1 1 0 1 0 1 1 0 1 0 1 0 0 0 1 1 0 1 1 0 1 0 0 1 1 0 1 0 1 0 1 1 0 1
[166] 0 1 0 0 0 0 0 0 0 0 0 1 0 0 1 1 0 1 1 1 0 0 0 1 0 1 1 0 0 0 1 1 0 1 0 1 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0
[221] 0 0 0 1 0 1 1 1 1 0 0 0 0 1 0 1 1 1 0 1 0 0 0 0 0 0 0 1 1 1 0 1 0 1 0 0 1 0 0 0 1 0 1 0 0 1 0 1 0 0 0 1 0 0 0 0
[276] 0 1 0 1 1 0 1 0 1 0 0 0 1 0 0 0 1 1 0 1 0 1 0 0 0 0 0 1 1 0 0 1 0 0 0 0 0 0 1 1 1 0 1 0 1 0 1 1 0 1 0 0 1 0 1 0 0
[331] 1 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 1 1 0 0 0 1 0 1 0 1 0 0 1 0 0 1 0 0 0 0 1 0 1
[386] 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 1 0 0 0 1 0 1 1 0 1 0 1 1 1 1 1 0 0 0 1 1 0 0 1 1 0 0 0 0 0 0 0 0 0
[441] 0 1 1 0 1 0 0 1 1 0
Levels: 0 1
> #karisiklik matrisi
> karisiklikmatrisi <- table(memekanseri[,3],ongoru)
> karisiklikmatrisi
    ongoru
       0   1
  1  210   4
  2   38   4
  3   30  11
  4    6  26
  5    0  19
  6    1  18
  7    0  21
  8    0  17
  9    0   6
  10   0  39
> #dogruluk oranini bulmamız lazım onemli konu
> sum(diag(karisiklikmatrisi))/sum(karisiklikmatrisi)
[1] 0.4755556
```

Karar Ağacı İkili Sınıflandırma Matrisi

		Gerçek	
		0	1
	0	279	11
Tahmin	1	6	154

Karar Ağacı İkili Sınıflandırma Algoritmaları Oranları

Karar Ağacı	Test Seti
Doğruluk Oranı	%0,96
Hata Oranı	%0,37
Kesinlik Oranı	%0,96
Duyarlılık Oranı	%0,97
Belirleyicilik Oranı	%0,93
F-Ölçüt	%0,97

Naive Bayes İkili Sınıflandırma Matrisi+

		Gerçek	
		0	1
	0	279	11
Tahmin	1	6	154

Naive Bayes İkili Sınıflandırma Algoritmaları Oranları

Karar Ağacı	Test Seti
Doğruluk Oranı	%0,96
Hata Oranı	%0,37
Kesinlik Oranı	%0,96
Duyarlılık Oranı	%0,97
Belirleyicilik Oranı	%0,93
F-Ölçüt	%0,97

1.1 Oluşturulan Modellerin Başarım Ölçütleri

	TP	FP	FN	TN	Doğruluk	Hata	Kesinlik	Duyarlılık	Belirleyicilik	F-Ölçütü
Karar Ağacı	279	11	6	154	0,96	%0,37	%0,96	%0,97	%0,93	%0,97
Naive Bayes	279	11	6	154	0,96	%0,37	%0,96	%0,97	%0,93	%0,97

KARAR AĞACI VE BAYES KIYASLAMASINI

Sonuç olarak ikisi de %50 şekliden çıkmıştır. Doğruluk , hata , kesinlik ,duyarlılık , belirleyicilik oranları eşittir. İki durumda kullanıma uygundur.

SONUÇ ; iki durumda kullanılabilir.

KAYNAKÇA

https://www.google.com/search?q=naive+bayes+nedir&bih=714&biw=1536&hl=tr&sxsrf=ALiCzsYmXiDhVwb-d_bFvPqEfW5kCeFEHQ%3A1672091309518&ei=rRaqY5qgH5-Mxc8Py7yPkAM&oq=Naive+Bayes+&gs_lcp=Cgxnd3Mtd2l6LXNlcnAQARgAMgQIABBDMgUIABCABDIFCAAQgAQyBAgAEEMyBQgAEIAEMgUIABCABDIFCAAQgAQyBQgAEIAEMgUIABCABDIFCAAQgAQ6CggAEEcQ1gQQsAM6BwgAELADEEM6BwgjEOoCECc6DAgAEOoCELQCEEMYAUoECEEYAEoECEYYAVCeGFieGGCtKGgEcAF4AIABbogBbpIBAzAuMZgBAKABAaABArABEMgBCsABAdoBBggBEAEYAQ&sclient=gws-wiz-serp

https://www.drozdogan.com/kanserin-sebepleri-kisaca-nelerdir/

https://www.google.com/search?q=karar+a%C4%9Fac%C4%B1+nedir&bih=714&biw=1536&hl=tr&sxsrf=ALiCzsbNNUJA5_JDClsoUKUu3FYhX577lw%3A1672090236960&ei=fBKqY9iVOoCExc8Pnpa0-AM&oq=karar+a%C4%9Fac%C4%B1+ne&gs_lcp=Cgxnd3Mtd2l6LXNlcnAQAxgAMgUIABCABDIFCAAQgAQyBggAEBYQHjIGCAAQFhAeMgYIABAWEB4yBggAEBYQHjIGCAAQFhAeMgYIABAWEB4yBggAEBYQHjIGCAAQFhAeOgcIIxDqAhAnOgQIIxAnOgQIABBDOgsIABCABBCxAxCDAToICAAQsQMQgwE6BAguEEM6CggAELEDEIMBEEM6DQgAEIAEELEDEIMBEAo6BwgAEIAEEAo6BQguEIAESgQIQRgASgQIRhgAUKEGWIIlYNguaAFwAXgAgAF0iAGiC5IBAzUuOZgBAKABAbABCsABAQ&sclient=gws-wiz-serp

https://www.saglik.org.tr/post/kadinlar-icin-saglik-kontrolleri?gclid=CjwKCAiAqaWdBhAvEiwAGAQlth2PghX6R2O-99GjhcQNaE5WmH8S7WXD-aKbLhJV5h9j8eZBCZW-hBoCgXEQAvD_BwE

https://www.memorial.com.tr/hastaliklar/meme-kanseri-belirtileri-tanisi-ve-tedavi-yontemleri

https://www.gurkanyetkin.com/meme-kistleri/

https://ders.bilecik.edu.tr/my/

.