T. C.

BİLECİK ŞEYH EDEBALİ ÜNİVERSİTESİ

İKTİSADİ VE İDARİ BİLİMLER FAKÜLTESİ

YÖNETİM BİLİŞİM SİSTEMLERİ

ŞARAP VERİ ANALİTİĞİ ANALİZİ

HAZIRLAYAN

MUSTAFA BAN

BİLECİK,2022

İÇİNDEKİLER

ÖNSÖZ

Alkol kalitesizliği günümüzde yaygın olarak görülen, geleceğe hastalık ve ölümcül sonuçlar yaratabilecek düzeyde öneme sahiptir. Bu sebepten dolayı bu çalışmamda konu olarak Şarap kalitesi seçilmiştir. Bir internet sitesinden alınan veri seti düzenlenerek, R programlama dili ile analiz edilmiştir.

Modelleme aşamasında K-en yakın komşu algoritması, Naive-Bayes sınıflandırma, C4.5 Karar ağaçları teknikleri kullanılmıştır. Elde edilen sonuçlar karşılaştırılmış ve yorumlanmıştır.

Bu çalışma sürecinde ve eğitim hayatıma katkıları, pozitif ve yapıcı tavırlarıyla desteklerini benden hiç esirgemeyen vizyoner, zarif ve değerli Hocam dr. Öğr. Üyesi Nur Kuban Torun'a teşekkür ederim.

MUSTAFA BAN

Bilecik, 2022

VERİ MADENCİLİĞİ

1.1Veri Madenciliği Nedir?

Veri madenciliği; önceden bilinmeyen, geçerli ve uygulanabilir bilginin veri yığınlarından dinamik bir süreç ile elde edilmesi olarak tanımlanabilir. Bu süreçte kümeleme, veri özetleme sınıflama kurallarının öğrenilmesi, bağımlılık ağlarının bulunması, değişkenlik analizi ve anomali tespiti gibi farklı birçok teknik kullanılmaktadır.

Veri madenciliği ile büyük veri yığınlarından oluşan database sistemleri içerisinde gizli kalmış bilgilerin çekilmesi sağlanır. Bu işlem, istatistik, matematik disiplinleri, modelleme teknikleri, database teknolojisi ve çeşitli bilgisayar programları kullanılarak yapılır. Veri madenciliği büyük miktarda veri inceleme amacı üzerine kurulmuş olduğu için veri tabanları ile yakından ilişkilidir. Gerekli verinin hızla ulaşılabilecek şekilde amaca uygun bir şekilde saklanması ve gerektiğinde hızla ulaşılabilmesi gerekir. Günümüzde yaygın olarak kullanılmaya başlanan veri ambarları günlük kullanılan veri tabanlarının birleştirilmiş ve işlemeye daha uygun bir özetini saklamayı amaçlar.

Veri madenciliği kendi başına bir çözüm değil çözüme ulaşmak için verilecek karar sürecini destekleyen, problemi çözmek için gerekli bilgileri sağlamaya yarayan bir araçtır. Veri madenciliği; analistin'e, iş yapma aşamasında oluşan veriler arasındaki şablonları ve ilişkileri bulması konusunda yardım etmektedir.

Şekil1:Veri madenciliği gelişim süreci

Gelişim Adımları	Cevaplanan Karar Problemi	Kullanılabilen Teknolojiler	Ürün Sağlayıcıları	Karakteristikler
Veri Toplama (1960'lar)	"Benim toplam karım geçen 5 yılda ne kadardı?"	Bilgisayarlar, Teypler, Diskler	IBM,CDC	Geriye dönük , statik veri dağıtımı
Veri Erişimi (1980'ler)	"İngiltere'de geçen mart ayında birim satışları ne kadardı?"	İlişkisel Veritabanları, SQL, ODBC	Oracle,Sybase, Informix,IBM, Microsoft	Kayıt düzeyinde geriye dönük, dinamik veri dağıtımı
Veri Ambarlama ve Karar Destek Sistemleri (1990'lar)	"İngiltere'de geçen mart ayında birim satışları ne kadardı?"	OLAP, Çok Boyutlu Veritabanı Sistemleri, Veri ambarları	Pilot, Comshare, Arbor,Cognos, Microstrategy	Çoklu düzeylerde, geriye dönük dinamik veri dağıtımı
Veri Madenciliği (Bugün)	"Gelecek ay Boston'daki birim satışlar muhtemelen ne olabilir, niçin?	İleri düzeyde algoritmalar, çok işlemcili bilgisayarlar, büyük veritabanları	Pilot, Lockheed, IBM,SGI, SPSS,SAS, Microsoft vs.	Geleceğe dönük ,proaktif enformasyon dağıtımı

1.2Veri Madenciliği Uygulama Alanları

- Veri tabanı analizi ve karar verme desteği

- Pazar Araştırması : Hedef pazar , müşteriler arası benzerliklerin saptanması, sepet analizi, çapraz pazar incelemesi

- Risk Analizi : Kalite kontrol, rekabet analizi, öngörü, sahtekarlıkların saptanması - Belgeler arası benzerlik : haber kümeleri, e-posta

- Müşteri kredi risk araştırmaları

- Kurum kaynaklarının en optimal biçimde kullanımı

- Geçmiş ve mevcut yapı analiz edilerek geleceğe yönelik tahminlerde bulunma

1.3 Pazarlama

– Müşterilerin satın alma örüntülerinin belirlenmesi,

- Müşterilerin demografik özellikleri arasındaki bağlantıların bulunması,

- Posta kampanyalarında cevap verme oranının artırılması,

- Mevcut müşterilerin elde tutulması, yeni müşterilerin kazanılması,

- Pazar sepeti analizi (Market Basket Analysis)

- Müşteri ilişkileri yönetimi (Customer Relationship Management)

- Müşteri değerlendirme (Customer Value Analysis)

- Satış tahmini (Sales Forecasting).

1.4 Bankacılık

- Farklı finansal göstergeler arasında gizli korelasyonların bulunması,
- Kredi kartı dolandırıcılıklarının tespiti,
- Kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi,
- Kredi taleplerinin değerlendirilmesi.

1.5 Sigortacılık

- Yeni poliçe talep edecek müşterilerin tahmin edilmesi,
- Sigorta dolandırıcılıklarının tespiti,
- Riskli müşteri örüntülerinin belirlenmesi

1.6 CRM(Müşteri İlişkileri Yönetimi)

-Müşteri sadakatinin artırılması.

-Pazarlama kampanyalarından en yüksek seviyede yarar sağlama çalışmalarının yapılması.

1.7 E-Ticaret:

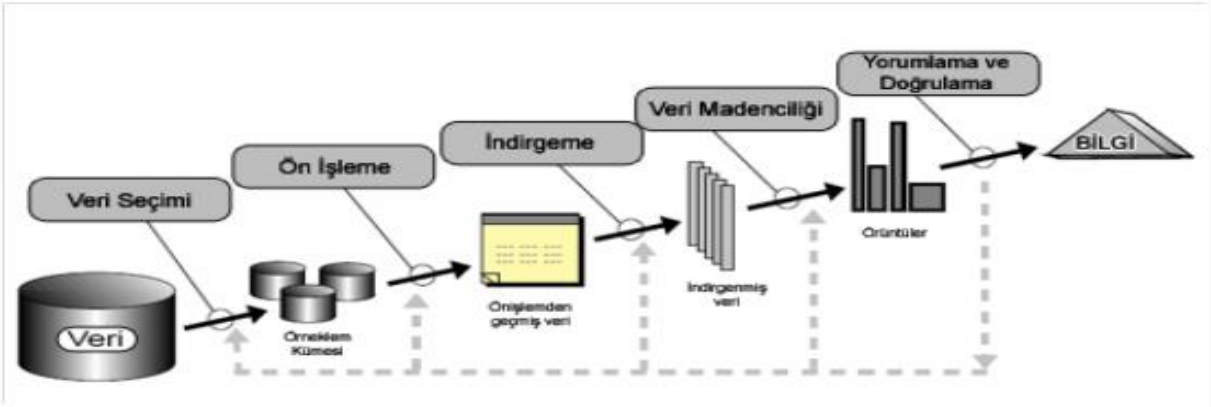
-Sunuculara yapılan saldırıların tespit edilmesi.

-Web sitesinde gezinen kullanıcıların davranışlarının belirlenmesi.

2. Veri Madenciliği Aşamaları

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 madenciliği yöntemlerinin uygulanabilmesi için veri ambarlarında veya veri tabanlarında tutulan verilerin belirli aşamalardan geçmesi gerekir. Başarılı bir veri madenciliği çalışmasında uygulanması gereken aşamalar aşağıda verilmiştir.

Şekil 2.Bilgi keşif sürecinde veri madenciliği



2.1 Veri Seçimi

Veri seçimi, veri madenciliği aşamalarında en fazla zaman alan kısımlardan biridir. Bu aşamada bilgi sistemlerinde oluşan bilgi iyi analiz edilmelidir ve problemle ilişkilendirilmelidir. Analizi yapan kişinin veri kalitesini ölçmesi açısından bu aşama önemlidir. Büyük miktardaki verilerin tek bir veri tabanı veya veri ambarında birleştirilmesi veri madenciliği uygulaması için gereklidir. Veri seçimi aşaması filtreleme olarak da isimlendirilebilir.

2.2 Ön İşleme

Ön işleme aşaması veri madenciliğinin başarısı için önemlidir. Bu aşamada veri, sonraki aşamalarda kullanılabilmesi için elverişli hale getirilir. Ön işleme aşamasının başarısı sonuçtaki başarıyı doğrudan etkiler. Başarılı bir ön işleme aşamasıyla kesin ve net sonuçlara ulaşmak mümkündür.

2.3 İndirgeme

Veri üzerinden faydalı ve doğru sonuç elde etmek için kullanılacak verinin indirgenmesi gerekir. Eldeki verinin büyük bir kısmı, her ne kadar ön işleme aşamasından geçmiş olsa bile sonraki aşamalarda kullanılabilecek durumda değildir. Dolayısıyla kullanılabilecek duruma indirgenmesi gerekir.

2.4 Veri Madenciliği

Veri madenciliği çalışmasının tam olarak kullanıldığı aşamadır. Veri bu aşamaya gelince doğru ve kullanılabilir haldedir. Çalışmanın amacına göre bu aşamada veri madenciliği yöntemlerinden biri veya birkaçı uygulanır. Gerektiği durumlarda farklı yöntemler birleştirilerek kullanılabilir.

2.5 Yorumlama ve Doğrulama

Veri üzerinde veri madenciliği uygulandıktan sonra alınan sonuçlar yorumlanır ve çalışmanın doğru sonuca ulaşıp ulaşmadığı araştırılır. Bu aşamada genellikle farklı yöntemler uygulanmışsa onların karşılaştırması yapılır. Elde edilen sonuçlar yapılmış olan diğer çalışmaların sonuçlarıyla karşılaştırılıp doğrulanır.

3. PROBLEMİN TANIMLANMASI

Şarap firması birçok şarap satışı gerçekleştiriyor bu şarapların kalitesini, kullanılabilir olmasına bakılacak. Veri setimizde 8 adet değişken vardır Bu değişkenler; Uçucu asitlik, Sitrik asit, Klorürler, Serbest kükürt dioksit, Toplam kükürt dioksit, Yoğunluk, Sülfatlar, Kalite olarak belirlenmiştir. Bu değişkenlerden yola çıkılarak Şarapların kalitesine içilebileceğine bakacağız.

3.1Veri Setini Anlama

Kullanılan veri seti Uci(data seti) sayfasından alınmıştır . Bu veri setinde Şarap içenlerin kimlik bilgileri bulunmamaktadır. Veri setinde şarap denemesi sonucunda 4'en 8'e kadar puan verilmiştir. Kalite dışında geriye kalan 7 değerim nümerik değerlerdir.

Veri setinde 7 nümerik değer vardır bunlar; Uçucu asitlik, Sitrik asit, Klorürler, Serbest kükürt dioksit, Toplam kükürt dioksit, Yoğunluk, Sülfatlardır. Kalite kısmında beğenilenler 4'ten 8'e kadar numaralandırılmıştır.

Tablo1:Veri setinde bulunan niteliklere ait özellikler

	TAHMİN İÇİN KULLANILAN VERİNİN YAPISI		
	Değişken	**Veri tipi**	**Veri setinde Gösterimi**
1	**Uçucu asitlik**	**Numerik**	
2	**Sitrik asit**	**Nümerik**	
3	**Klorürler**	**Nümerik**	
4	**Serbest kükürt dioksit**	**Nümerik**	
5	**Toplam kükürt dioksit**	**Nümerik**	
6	**Yoğunluk**	**Nümerik**	
7	**Sülfatlar**	**Nümerik**	
8	**Kalite**	**Factor**	**4,5,6,7,8,**

3.2Analize Hazırlık

Bundan sonraki aşamalar R Studio'da yapılmıştır. Uygulama kodları eklerdedir.

Veri seti 239 gözlem ve 8 değişkenden oluşmaktadır.

Değişkenler sırasıyla Uçucu asitlik, Sitrik asit, Klorürler, Serbest kükürt dioksit, Toplam kükürt dioksit, Yoğunluk, Sülfatlar, Kalite şeklindedir. Öncelikle veri setinin yapısı incelenmiş, nümerik ve faktör şeklinde düzenlenmiştir. Nümerik değişkenler nümerik olarak diğer değişkenlerde faktör şeklinde tanımlanmıştır. Düzenlendikten sonra şu hale dönüşmüştür.

$ Uçucu.asitlik : num 0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...

$ Sitrik.asit : num 0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...

$ Klorürler : num 0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...

$ Serbest.kükürt.dioksit: num 45 14 30 47 47 30 30 45 14 28 ...

$ Toplam.kükürt.dioksit : num 170 132 97 186 186 97 136 170 132 129 ...

$ Yoğunluk : num 1.001 0.994 0.995 0.996 0.996 ...

$ Sülfatlar : num 0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...

$ Kalite : Factor w/ 5 levels "4","5","6","7",..: 3 3 3 3 3 3 3 3 3 3 ...

Veri setinin özeti Tablo 2'de verilmiştir. Bu tabloda kategorik değişkenler ve nümerik değişkenlerin minimum değerleri, 1. kartil, medyan, ortalama, 3. kartil ve maksimum değerleri görülür.

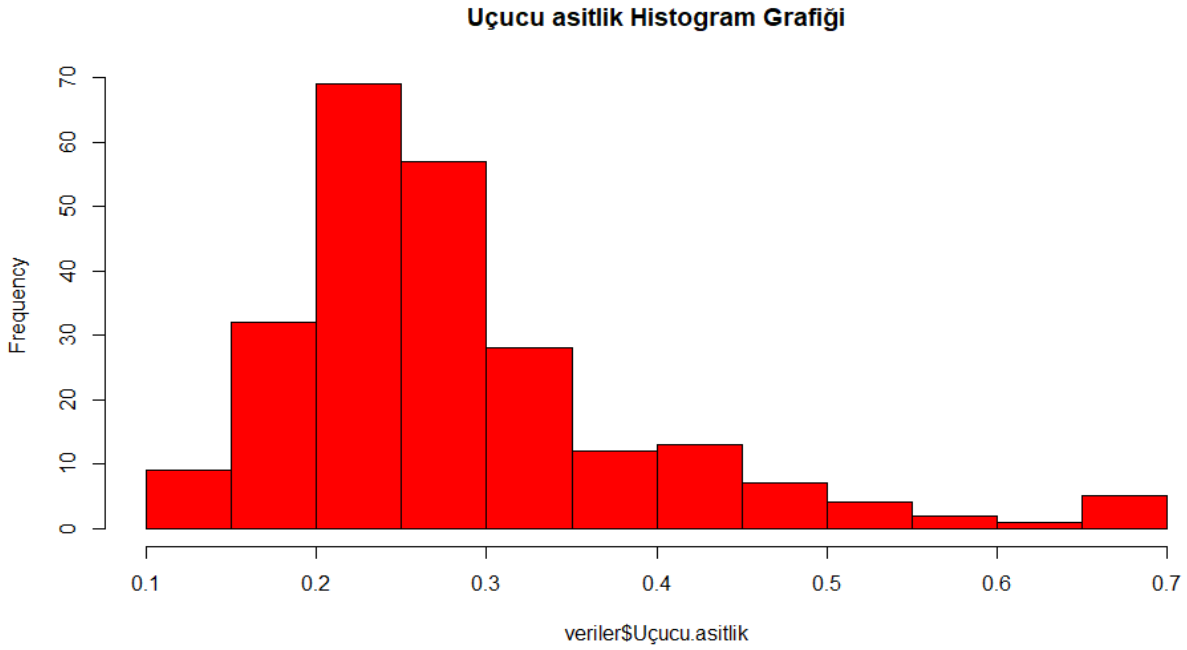
Tablo 2:Veri seti özeti

Uçucu asitlik	**Sitrik asit**	**Klorürler**	**Serbest kükürt dioksit**	**Toplam kükürt dioksit**	**Yoğunluk**	**Sülfatlar**	**Kalite**
Min. :0.1200	Min. :0.0000	Min. :0.02000	Min. :4.00	Min. :47.0	Min. :0.9892	Min. :0.2700	11
1st Qu.:0.2300	1st Qu.:0.2800	1st Qu.:0.04000	1st Qu.:27.50	1st Qu.:115.0	1st Qu.:0.9926	1st Qu.:0.3900	88
Median :0.2600	Median :0.3400	Median :0.04600	Median :38.00	Median :150.0	Median :0.9949	Median :0.4600	104

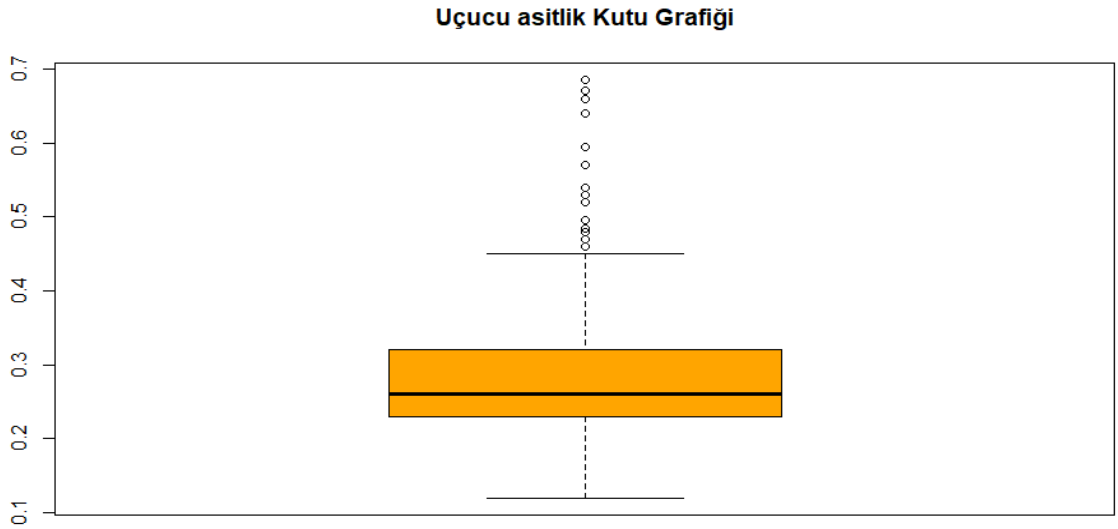
Mean :0.2871	Mean :0.3482	Mean :0.05097	Mean :38.76	Mean :148.2	Mean :1.0324	Mean :0.4661	28
3rd Qu.:0.3200	3rd Qu.:0.4000	3rd Qu.:0.05400	3rd Qu.:50.75	3rd Qu.:178.0	3rd Qu.:0.9972	3rd Qu.:0.5200	8
Max. :0.6850	Max. :0.8800	Max. :0.20000	Max. :81.00	Max. :272.0	Max. :10.0020	Max. :0.8400	25

Veri setindeki değişkenler tek tek incelenmiştir. Bunun için her birine uygun grafikler çizilmiştir. Nümerik değişkenler için histogram grafikleri, kategorik değişkenler için çubuk grafikleri çizilmiştir. Ayrıca değişkenler kutu grafikleri ile de gösterilmiştir. Böylece dağılımları hakkında daha kolay bilgi edinilmiştir.

Şekil 3: Uçucu asitlik değişken grafiği

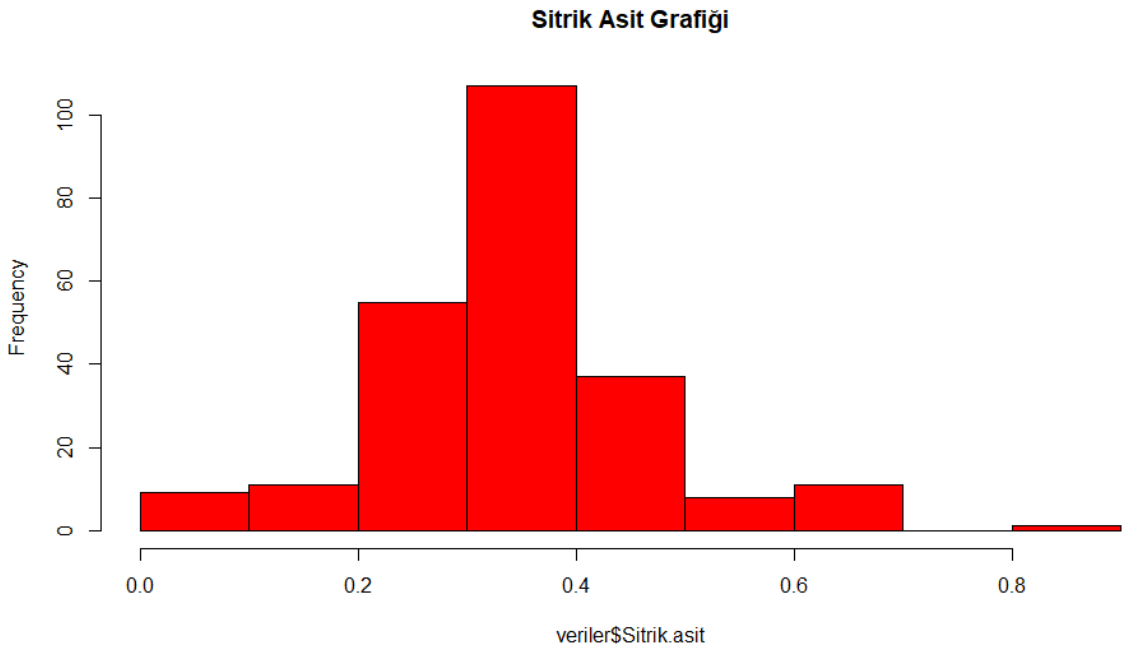


Şekil 4:Uçucu asitlik değişken kutu grafiği

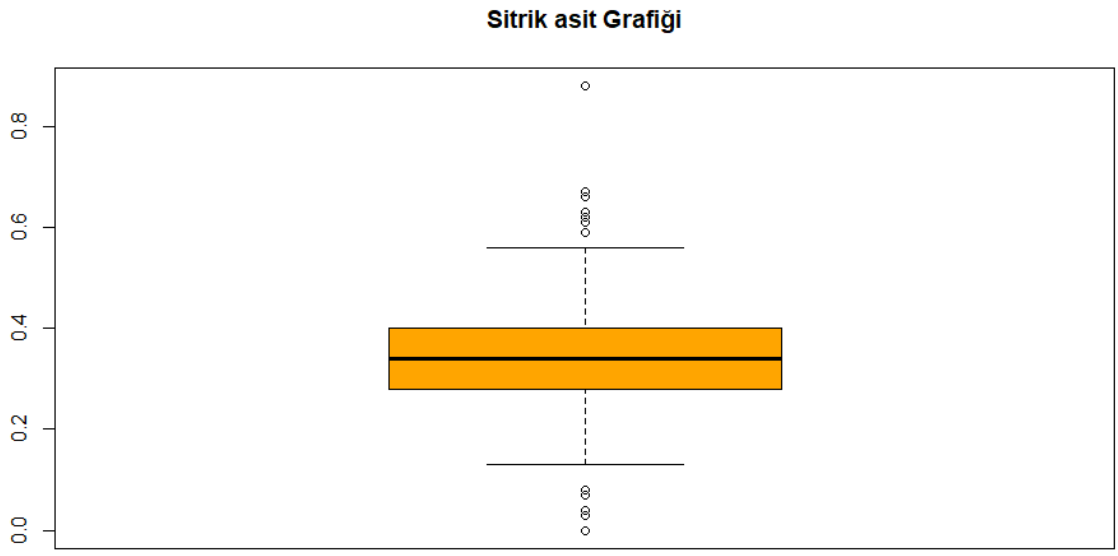


Veri setinde Uçucu asitlik değişim aralığı en küçük 0 büyük ise 0.7 arasındadır. Frekanslara bakıldığında 0.2 civarında uçucu asitliğin fazla olduğu görülmektedir.

Şekil 5:Sitrik asit değişken grafiği

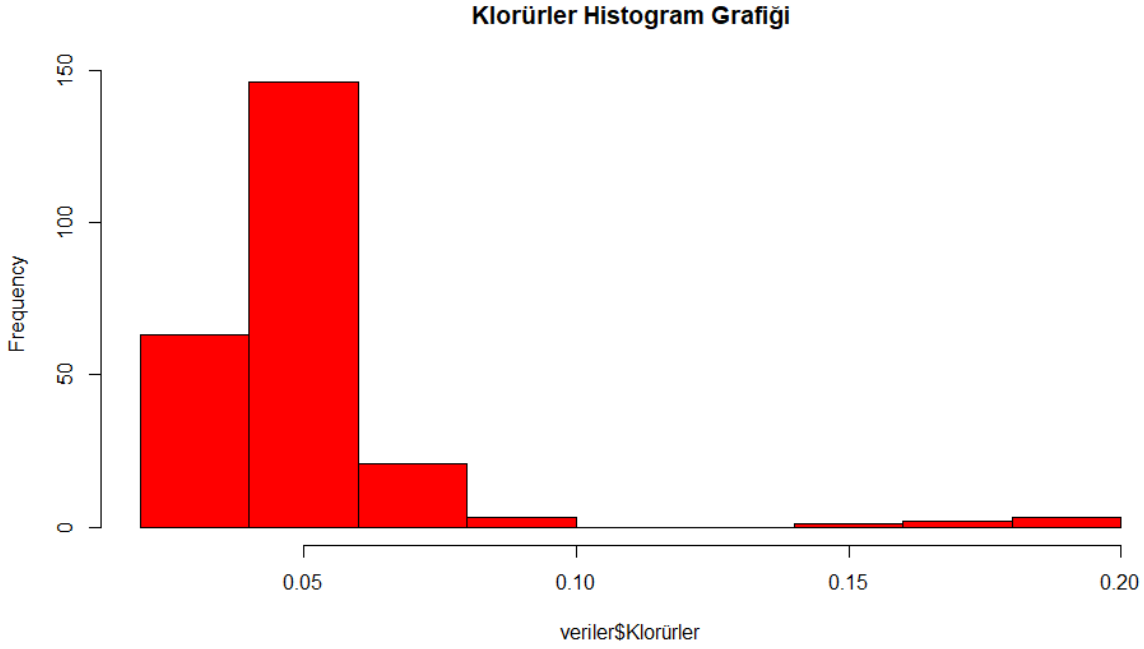


Şekil 6:Sitrik asit değişken kutu grafiği

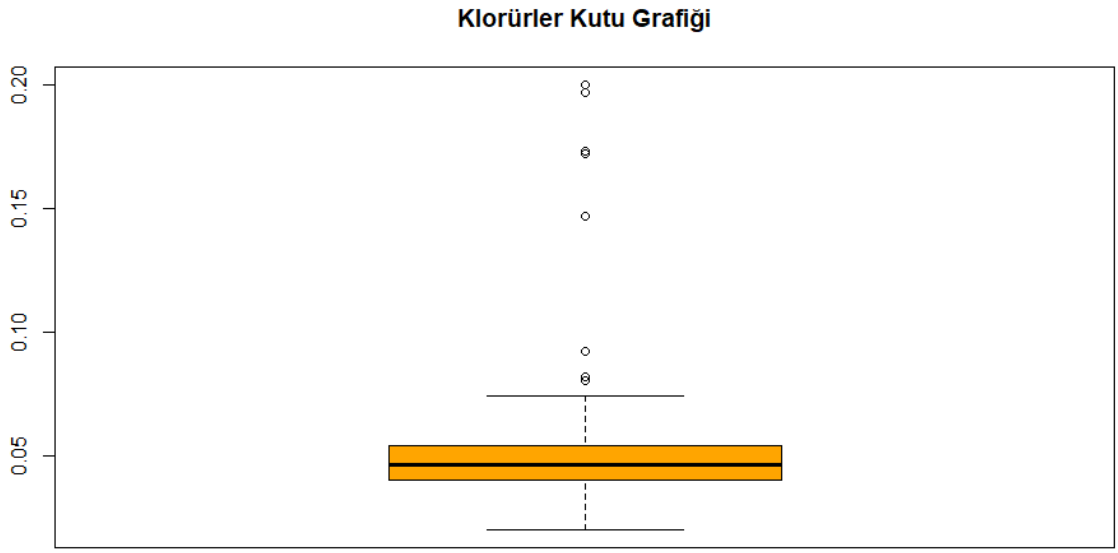


Sitrik asit değişken değeri 0 ile 0.8 arasındadır. Yoğunluk olarak ise 0.4’dir

Şekil 7:Klorürler değişken grafiği

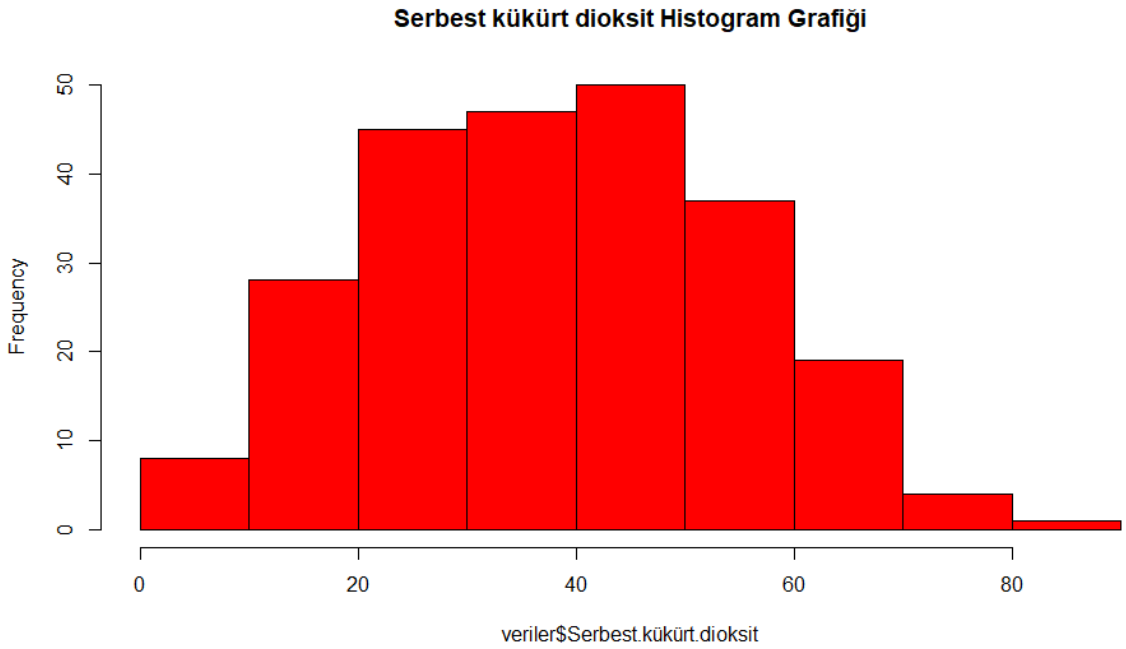


Şekil 8:Klorürler değişken kutu grafiği

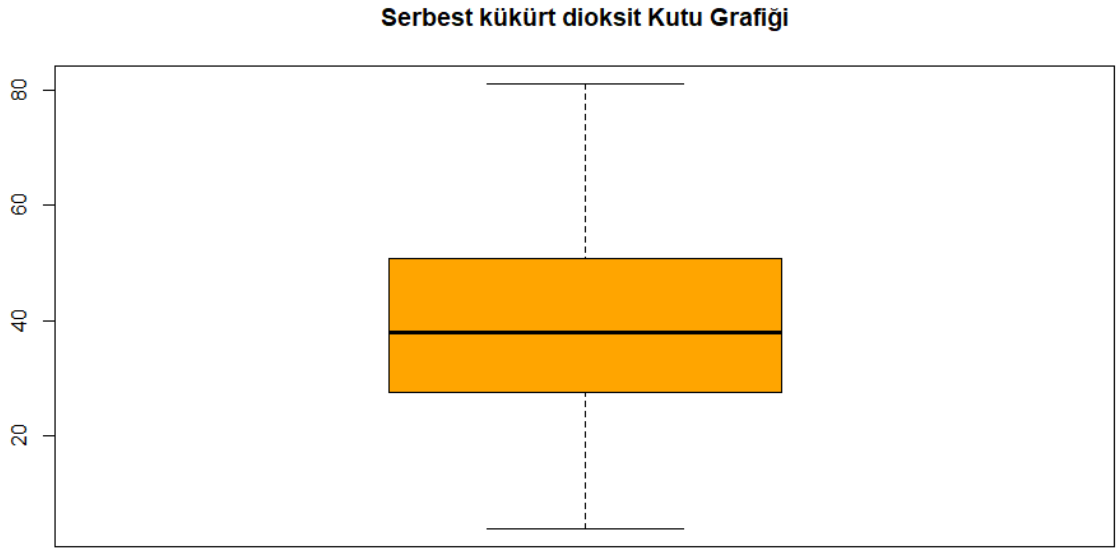


Klorür değişkeni değerleri 0 ile 0.20 arasında değişmektedir. En fazla tekrar eden değişkenler ise 0.05 ile 0.10 arasında görülmektedir.

Şekil 9:Serbest kükürt dioksit değişken grafiği

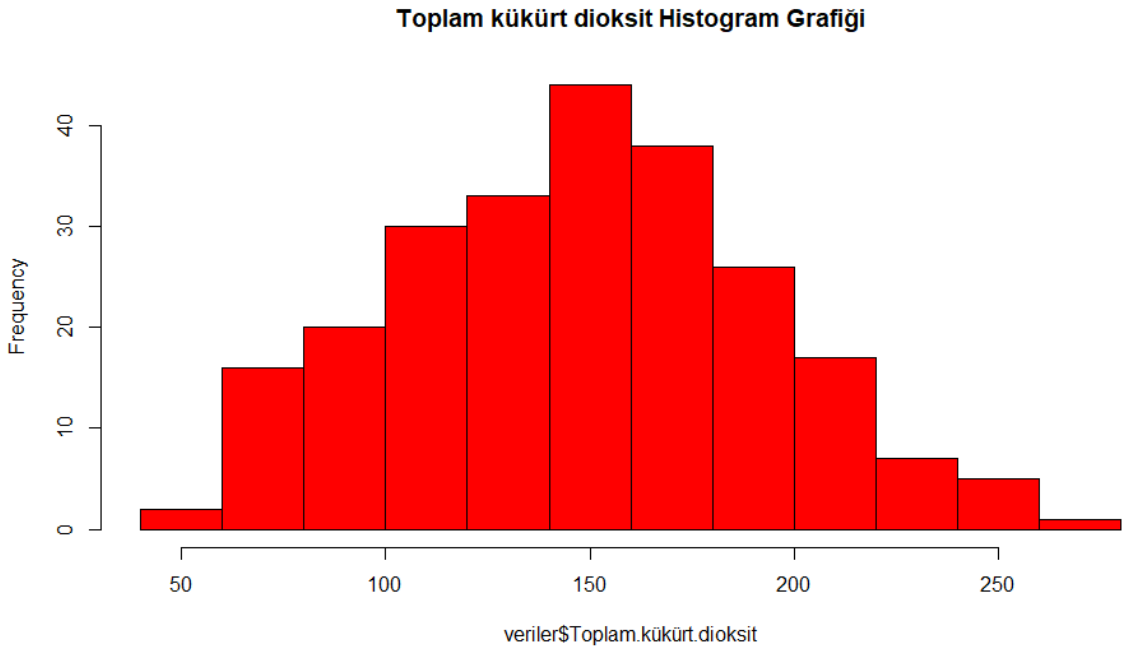


Şekil 10:Serbest kükürt dioksit değişken kutu grafiği

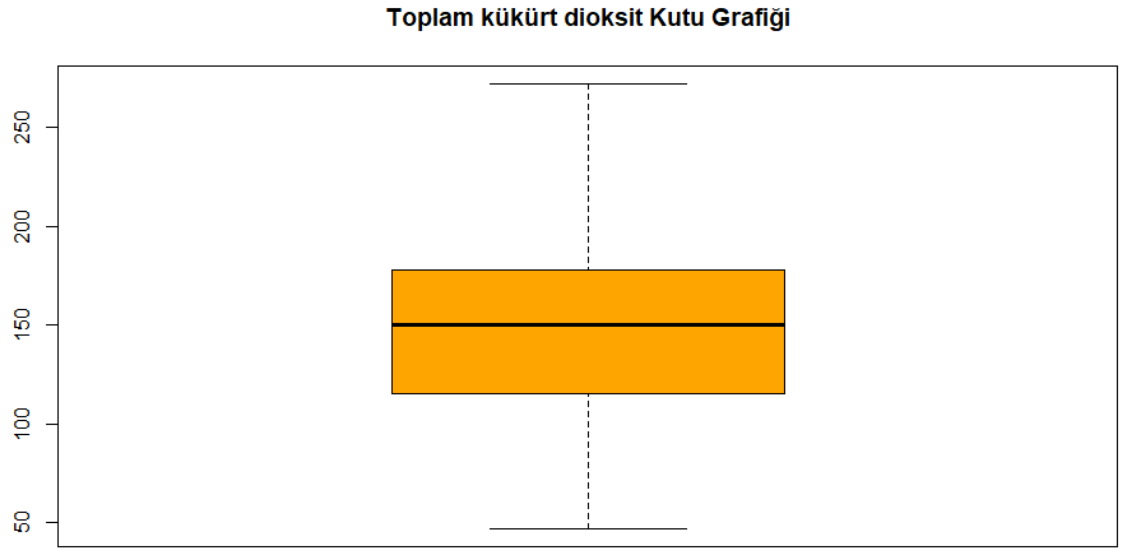


Serbest kükürt değişkeni değerleri 0 ile 80 arasındadır. 30 ile 50 arasında ise sık tekrar eden değişkenler vardır.

Şekil 11: Toplam kükürt dioksit değişken grafiği

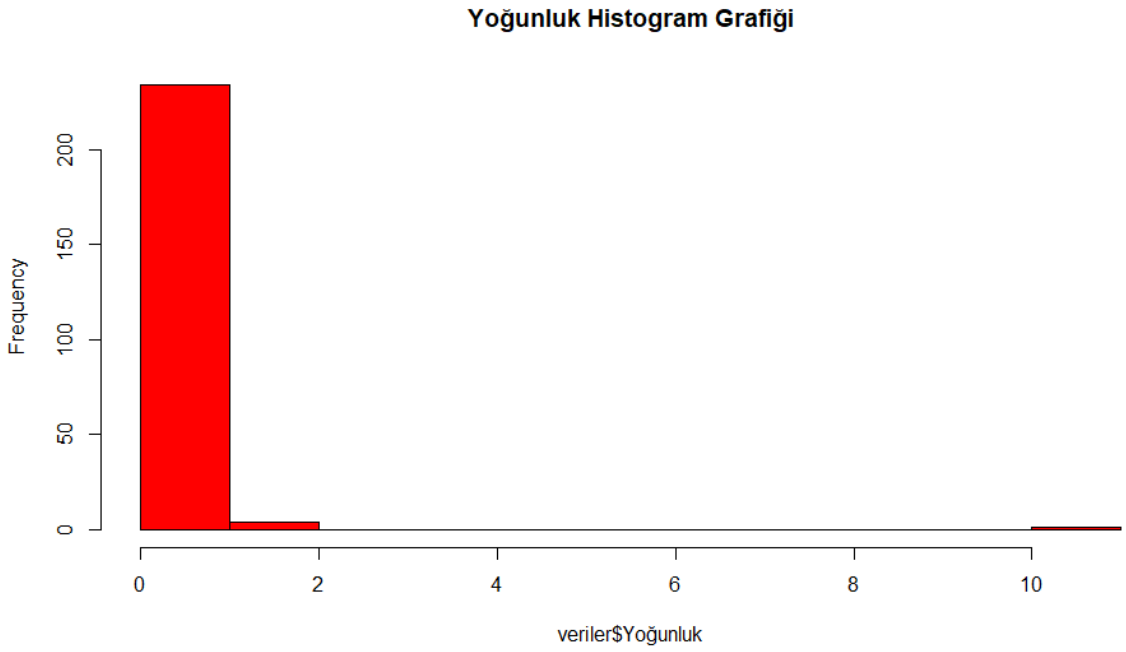


Şekil 12: Toplam kükürt dioksit değişken kutu grafiği

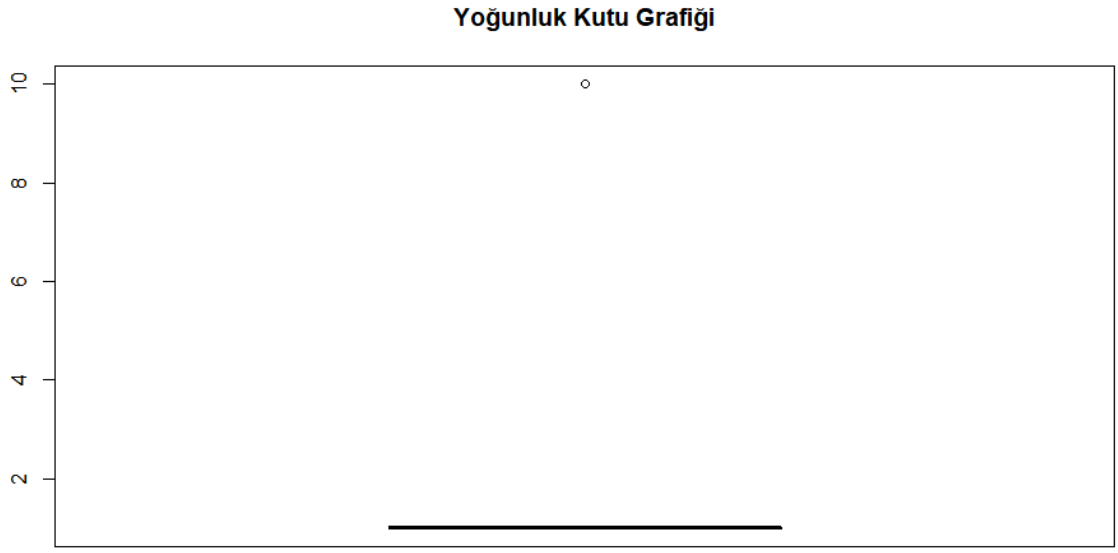


Toplam kükürt değişkeni değerleri 20 ile 300 arasında değişmektedir. Ortalaması 160’dır. Sık tekrar edilen değeler 180 ile 280 arasındadır.

Şekil 13: Yoğunluk değişken grafiği

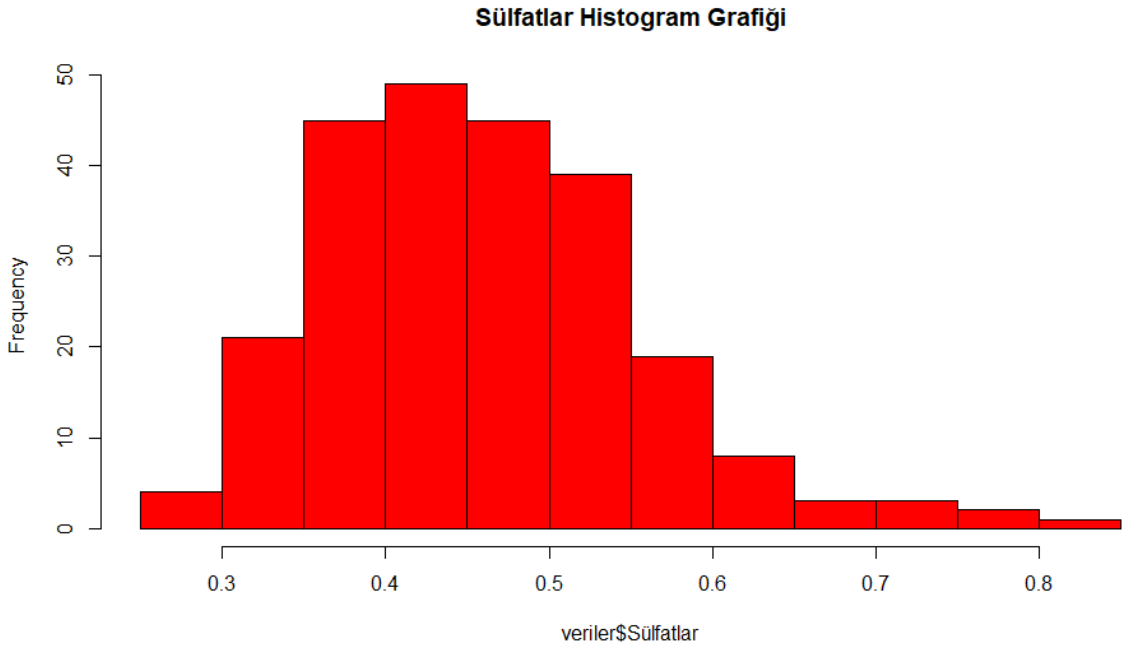


Şekil 14: Yoğunluk değişken kutu grafiği

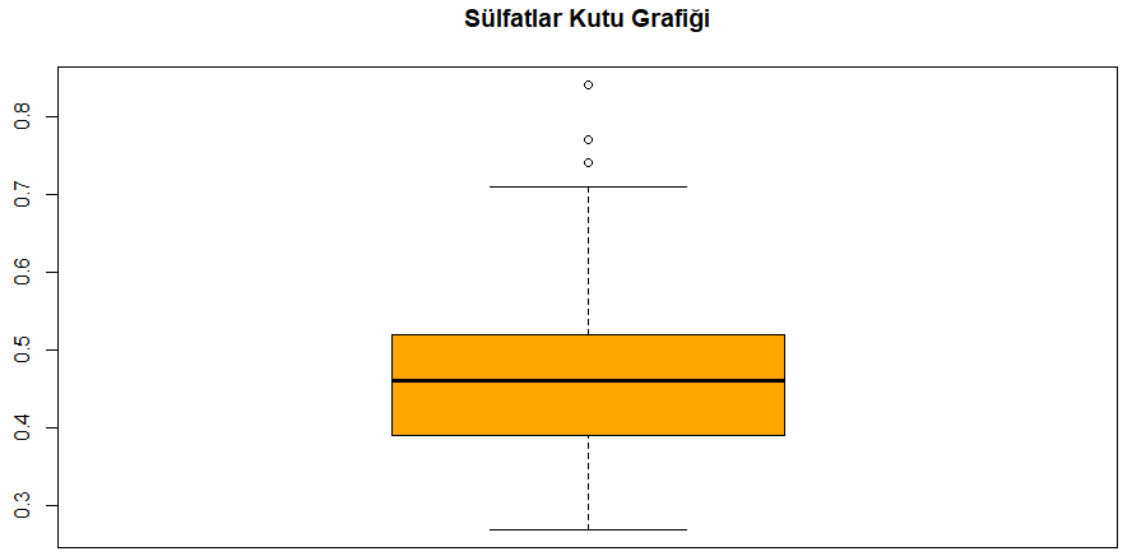


Yoğunluk değişkeni değerleri 0 ile 11 arasındadır. En çok tekrar eden değerleri 0 ile 1 arasında olduğu görülmektedir.

Şekil 15: Sülfatlar değişken grafiği

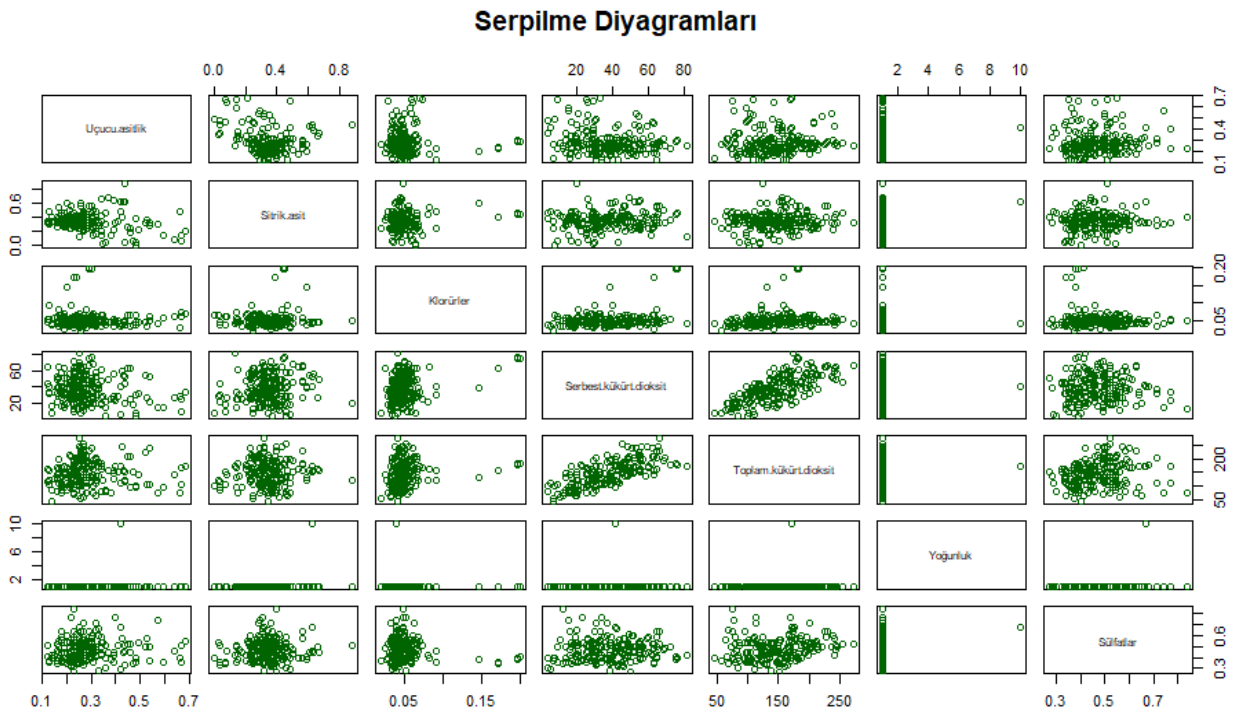


Şekil 16: Sülfatlar değişken kutu grafiği



Sülfatlar değişkeni değerleri 0 ile 0.9 arasında değişmektedir. En çok tekrar eden değerleri 0.4 ile 0.5 arasındadır.

Şekil 16: Serpilme Diyagramları



3.3 KNN Algoritması

K En Yakın Komşu (kNN) algoritması, ilk olarak 1950'lerin başında ortaya atılmıştır. Bu yöntemde, büyük eğitim setleri verildiğinde öğrenme işleminde oldukça zaman kaybedilmektedir. Bu nedenle, bilgi işlem gücü kullanılabilir hale gelene kadar popülerlik kazanmamıştır. 1960'lardan sonra ise, 1965'te N.J. Nilsson tarafından hazırlanan minimum uzaklık sınıflayıcı üzerine çalışmalarla geliştirilmiş; 1967'de T. Cover ve P. Hart'ın sunduğu "Yakın Komşular Örüntü Sınıflama" çalışmalarıyla netlik kazanmıştır.

Denetimli öğrenme yöntemlerinden biri olan k En Yakın Komşu algoritması hem sınıflama hem de regresyon ayağında kullanılabilen çok yönlü bir algoritmadır. En basit haliyle tanımlayacak olursak, sınıfı bilinmeyen veri, eğitim setindeki diğer veriler ile karşılaştırılır ve bir uzaklık ölçümü yapılır. Hesaplanan uzaklığa göre henüz bir sınıfa atanamamış veriye en optimal sınıf bulunur.

Hemen hemen her sınıflandırma modeli kendi içinde bir artık sınıflayıcı oluşturur ve gelen her yeni veride bu sınıflayıcı kullanılır. kNN algoritmasında ise bu tür bir artık sınıflayıcı

bulunmaz, bunun yerine gelen her yeni örnek için en yakın komşu kümesi tekrardan aranır. En Yakın Komşu sınıflandırma yönteminde, önceden hiçbir sınıflandırıcı model oluşturulmadığı ve her yeni verinin sınıflandırılmasında ham eğitim verilerine geri dönüldüğünden, eğitim kümesi tamamı sınıflandırıcı olarak değerlendirilir. Bu özelliği bakımından tembel öğrenici olarak nitelendirilen k En Yakın Komşu algoritması, her bir örnekte tek tek tarama yaptığı için sınıflama süreci uzun olan bir algoritmadır kNN algoritması, yeni verilerin hızla geldiği ve eğitim kümesinin hızla değiştiği durumlarda diğer algoritmalara göre daha iyi bir sınıflandırıcı olarak değerlendirilebilir. kNN algoritmasında en önemli hususlardan biri optimal k sınıf değerini bulmaktır. Sınıf değeri k, önceden belirlenir. En uygun k değeri verilerin boyutuna ve yapısına bağlıdır; k=1'den gözlem sayısı n'e kadar sınıf yaratmak mümkündür. Sınıf değerini olması gerekenden büyük kullanmak, çok benzer olmayan verileri aynı gruba alacağından, sınıflamada doğruluk değerini aşağı çekecektir. Tam tersi gereğinden küçük bir k değeri kullanmak ise bazı olası sınıfları saf dışı bırakacaktır; bu durumda yine sınıf doğruluğu aşağı yönlü ivme kazanır.

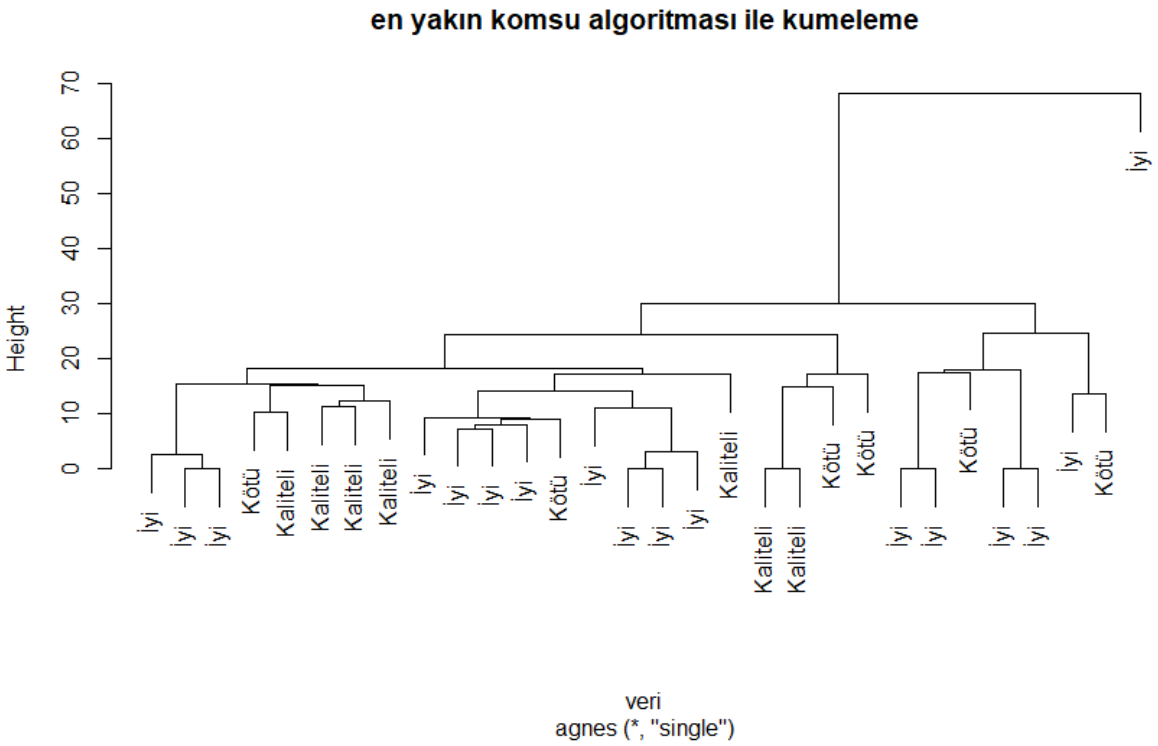
Bu çalışmada Öklid uzaklığı ile uzaklıklar hesaplanarak K-en yakın komşu algoritması uygulanmıştır. Bunun için, veri setinden yalnızca nümerik değerler taşıyan ve hedef niteliğin de olduğu bir alt küme elde edilmiştir. Bu değişkenler, "Uçucu asitlik, Sitrik asit, Klorürler, Serbest kükürt dioksit, Toplam kükürt dioksit, Yoğunluk, Sülfatlardır" değişkenleridir. Hedef nitelik de "Kalite" değişkenidir.

Formülü uyguladığımızda 239 veri ve 8 değişkenden oluşan yeni bir data.frame elde edilir.

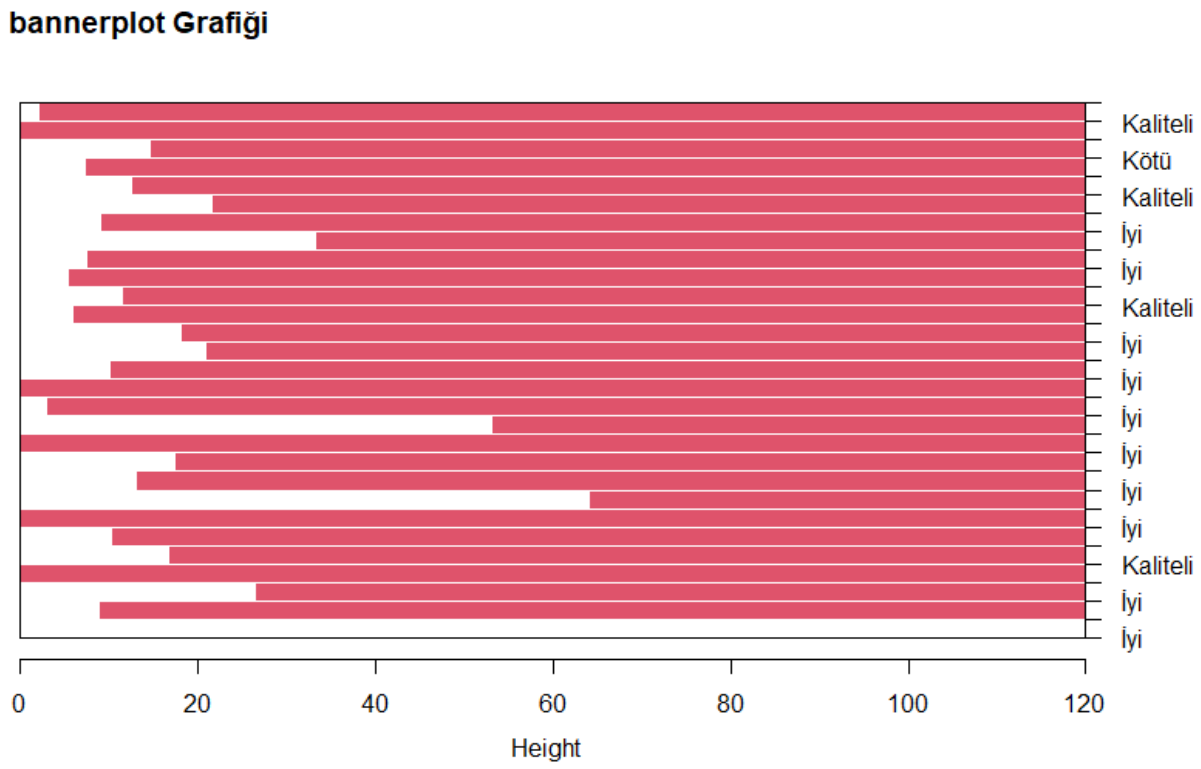
Modelin performansının ölçülmesi için kontenjans tablosu kurulur.

KNN algoritmasında "Şarap Kalite Kontrolü" incelenmiştir. En yakın Komşu Algoritması ve Bannerpot Grafiği çıkarılmıştır.

Şekil 17: En yakın Komşu algoritması ile kümeleme



Şekil 18: Bannerplot Grafiği



Bu çalışmada gerçek veriler ile Şarap kalitesi yapılmaya çalışıldığından Şaraplar 4 ile 8 arasında puanlandırılmıştır. 4-5 arası Şarap kalitesi “Kötü”, 6 Şarap kalitesi “İyi” ve 7-8 arası Şarap Kalitesi “Kaliteli” olarak belirlenmiştir. Veri setinden rast gele çağırdığım 30 kişiyi En yakın komşu algoritmasında ağaçlandırma yoluyla Türlerine ayırarak kendi içlerinde gruplanma yaptım ve 2 farklı Grafik oluşturdum. Veri setinde bulunan 239 kişi içerisinden rast gele 30 kişi çağırılmıştır.

3.4 C4.5 (KARAR AĞACI) Algoritması

C4.5 Algoritması bir karar ağacı sınıflandırma algoritmasıdır. Bu çalışmada J. Ross Quinlan tarafından geliştirilmiş C4.5 sınıflandırma algoritması üzerinde yeni bir budama algoritması önerilecektir. C4.5 iki işlem adımı ile gerçekleştirilmektedir. Bunlardan ilki ağacı oluşturma işlemi ve diğeri ise budama işlemidir.

Bu çalışmada budama işlemi için bir algoritma önerilmiştir. Weka yazılım ortamında C4.5 algoritması için Güven Faktörü (CF) parametresi kullanılmaktadır. Bu parametre ağacın gelişiminden sonra ağacın budaması işleminin derecesini belirlerken değeri de 0 ile 1 arasında değişmektedir. CF katsayısını azaltmak; son budamanın küçülmesine, CF katsayısı arttırmak ise; son budamanın büyümesine sebep olur. Önerilen metoda göre ilk önce veri seti niteliklerine göre C4.5 karar ağacı kurulmuştur. Son budama işlemi için Genetik Algoritma ile CF katsayısı belirlenmiş ve ağaç seçilen CF katsayısı ile budanmıştır. Budama işleminden sonra 10 katlı çapraz doğrulama işlemi ile sınıflandırma işlemi doğruluk oranları hesaplanmıştır En büyük bilgi kazancını sağlayacak biçimde bir eşik değer belirlenir. Eşik değeri belirlemek için tüm değerler sıralanır ve ikiye bölünür. Nitelikteki değerler eşik değere göre iki kategoriye ayrılmış olur. C4.5 karar ağacı algoritması uygulanmadan önce veri setinin yapısı incelenmiştir. Değişkenler nümerik ve faktör şeklinde atanmıştır. Sınıflandırma algoritması olduğu için veri seti eğitim ve test veri seti olarak ayrılmıştır.

Uygulamanın yapılabilmesi için R programlamaya RWeka paketi yüklenmiş ve kütüphaneden çağırılmıştır. Paketin içindeki J48() fonksiyonu C4.5 karar ağacı algoritması çözümünde kullanılmıştır.

Şarap veri setine uygulanan karar ağacı algoritması sonuçları verilmiştir.

Şekil 19: C4.5 Summary

```
=== Summary ===

Correctly Classified Instances         221               92.4686 %
Kappa statistic                          0.8779
Mean absolute error                      0.078
Root mean squared error                  0.1975
Relative absolute error                 18.9649 %
Root relative squared error             43.5704 %
Total Number of Instances              239
```

Burada correctly classified instances doğru yerleşen tahmin sayısıdır. Bunun toplam 239 kişi içerisinden 221 kişi olduğu gözükmektedir ve %92.4 doğruluk oranına sahiptir.

Modelin oluşturduğu ağaç şu şekildedir:

Şekil 20: C4.5 J48 Pruned tree

```
J48 pruned tree
------------------

Yoğunluk <= 0.9918
|   Serbest.kükürt.dioksit <= 23
|   |   Yoğunluk <= 0.991: Kötü (6.0/1.0)
|   |   Yoğunluk > 0.991
|   |   |   Sitrik.asit <= 0.39: Kaliteli (4.0)
|   |   |   Sitrik.asit > 0.39: iyi (4.0)
|   Serbest.kükürt.dioksit > 23
|   |   Klorürler <= 0.038: Kaliteli (19.0/1.0)
|   |   Klorürler > 0.038
|   |   |   Yoğunluk <= 0.9903: Kaliteli (2.0)
|   |   |   Yoğunluk > 0.9903
|   |   |   |   Serbest.kükürt.dioksit <= 38: iyi (6.0)
|   |   |   |   Serbest.kükürt.dioksit > 38
|   |   |   |   |   Sülfatlar <= 0.4: iyi (2.0)
|   |   |   |   |   Sülfatlar > 0.4: Kaliteli (3.0)
Yoğunluk > 0.9918
|   Sitrik.asit <= 0.25
|   |   Serbest.kükürt.dioksit <= 33
|   |   |   Serbest.kükürt.dioksit <= 29
|   |   |   |   Yoğunluk <= 0.9932: iyi (3.0/1.0)
|   |   |   |   Yoğunluk > 0.9932: Kötü (11.0)
|   |   |   Serbest.kükürt.dioksit > 29: iyi (5.0)
|   |   Serbest.kükürt.dioksit > 33: Kötü (16.0)
|   Sitrik.asit > 0.25
|   |   Yoğunluk <= 0.9959
|   |   |   Uçucu.asitlik <= 0.34
|   |   |   |   Sülfatlar <= 0.52
|   |   |   |   |   Sülfatlar <= 0.35
|   |   |   |   |   |   Yoğunluk <= 0.9937: iyi (4.0/1.0)
|   |   |   |   |   |   Yoğunluk > 0.9937: Kötü (6.0)
|   |   |   |   |   Sülfatlar > 0.35: iyi (42.0/9.0)
|   |   |   |   Sülfatlar > 0.52
|   |   |   |   |   Serbest.kükürt.dioksit <= 18: Kötü (2.0)
|   |   |   |   |   Serbest.kükürt.dioksit > 18
|   |   |   |   |   |   Toplam.kükürt.dioksit <= 184
|   |   |   |   |   |   |   Yoğunluk <= 0.9932
|   |   |   |   |   |   |   |   Klorürler <= 0.041: iyi (2.0)
|   |   |   |   |   |   |   |   Klorürler > 0.041: Kaliteli (6.0/2.0)
|   |   |   |   |   |   |   Yoğunluk > 0.9932: iyi (7.0)
|   |   |   |   |   |   Toplam.kükürt.dioksit > 184: Kaliteli (3.0/1.0)
|   |   |   |   |   Toplam.kükürt.dioksit > 184: Kaliteli (3.0/1.0)
|   |   |   |   Uçucu.asitlik > 0.34: Kötü (7.0/1.0)
|   |   Yoğunluk > 0.9959
|   |   |   Yoğunluk <= 0.9979
|   |   |   |   Yoğunluk <= 0.9975
|   |   |   |   |   Klorürler <= 0.044
|   |   |   |   |   |   Klorürler <= 0.043
|   |   |   |   |   |   |   Sitrik.asit <= 0.42: iyi (2.0)
|   |   |   |   |   |   |   Sitrik.asit > 0.42: Kötü (3.0)
|   |   |   |   |   |   Klorürler > 0.043: iyi (7.0)
|   |   |   |   |   Klorürler > 0.044
|   |   |   |   |   |   Uçucu.asitlik <= 0.28
|   |   |   |   |   |   |   Sitrik.asit <= 0.38
|   |   |   |   |   |   |   |   Klorürler <= 0.053
|   |   |   |   |   |   |   |   |   Uçucu.asitlik <= 0.22: iyi (4.0)
|   |   |   |   |   |   |   |   |   Uçucu.asitlik > 0.22
|   |   |   |   |   |   |   |   |   |   Uçucu.asitlik <= 0.245: Kötü (4.0)
|   |   |   |   |   |   |   |   |   |   Uçucu.asitlik > 0.245: iyi (3.0/1.0)
|   |   |   |   |   |   |   |   Klorürler > 0.053: Kötü (7.0)
|   |   |   |   |   |   |   Sitrik.asit > 0.38: iyi (4.0)
|   |   |   |   |   |   Uçucu.asitlik > 0.28: Kötü (9.0)
|   |   |   |   Yoğunluk > 0.9975
|   |   |   |   |   Klorürler <= 0.038: Kötü (2.0)
|   |   |   |   |   Klorürler > 0.038: iyi (12.0)
|   |   |   Yoğunluk > 0.9979
|   |   |   |   Sitrik.asit <= 0.37
|   |   |   |   |   Sitrik.asit <= 0.33: Kötü (4.0)
|   |   |   |   |   Sitrik.asit > 0.33: iyi (5.0)
|   |   |   Sitrik.asit > 0.37: Kötü (13.0)

Number of Leaves  :     35

Size of the tree :      69
```

KURAL1: Yoğunluk <= 0.9918, Serbest.kükürt.dioksit <= 23, Yoğunluk <= 0.991: Şarap Kötü

Kural2: Yoğunluk > 0.991, Sitrik.asit <= 0.39: Şarap Kaliteli

Kural3: : Yoğunluk > 0.991, Sitrik.asit > 0.39:Şarap iyi

Kural4: Serbest.kükürt.dioksit > 23, Klorürler <= 0.038:Şarap Kaliteli

Kural5: Klorürler > 0.038, Yoğunluk <= 0.9903: Şarap Kaliteli

Kural6: Yoğunluk > 0.9903, Serbest.kükürt.dioksit <= 38: Şarap iyi

Kural7: Serbest.kükürt.dioksit > 38, Sülfatlar <= 0.4: Şarap iyi

Kural8: Serbest.kükürt.dioksit > 38, Sülfatlar > 0.4: Şarap Kaliteli

Kural9: Yoğunluk > 0.9918, Sitrik.asit <= 0.25, Serbest.kükürt.dioksit <= 33, Serbest.kükürt.dioksit <= 29, Yoğunluk <= 0.9932: Şarap iyi

Kural10: Yoğunluk > 0.9918, Sitrik.asit <= 0.25, Serbest.kükürt.dioksit <= 33, Serbest.kükürt.dioksit <= 29 , Yoğunluk > 0.9932: Şarap Kötü

Kural11: Serbest.kükürt.dioksit <= 29, Serbest.kükürt.dioksit > 29: Şarap iyi

Kural12: Serbest.kükürt.dioksit <= 33, Serbest.kükürt.dioksit > 33: Şarap Kötü

Kural13: Sitrik.asit > 0.25, Yoğunluk <= 0.9959, Uçucu.asitlik <= 0.34, Sülfatlar <= 0.52, Sülfatlar <= 0.35, Yoğunluk <= 0.9937: Şarap iyi

Kural14: Sitrik.asit > 0.25, Yoğunluk <= 0.9959, Uçucu.asitlik <= 0.34, Sülfatlar <= 0.52, Sülfatlar <= 0.35, Yoğunluk > 0.9937:Şarap Kötü

Kural15: Sülfatlar <= 0.35, Sülfatlar > 0.35: Şarap iyi

Kural17: Sülfatlar > 0.52, Serbest.kükürt.dioksit <= 18: Şarap Kötü

Kural18: Serbest.kükürt.dioksit > 18, Toplam.kükürt.dioksit <= 184, Yoğunluk <= 0.9932, Klorürler <= 0.041: Şarap iyi

Kural19: Serbest.kükürt.dioksit > 18, Toplam.kükürt.dioksit <= 184, Yoğunluk <= 0.9932, Klorürler > 0.041: Şarap Kaliteli

Kural20: Yoğunluk <= 0.9932, Yoğunluk > 0.9932:Şarap iyi

Kural21: Toplam.kükürt.dioksit <= 184, Toplam.kükürt.dioksit > 184: Şarap Kaliteli

Kural22: Uçucu.asitlik <= 0.34, Uçucu.asitlik > 0.34: Şarap Kötü

Kural23: Yoğunluk > 0.9959, Yoğunluk <= 0.9979, Yoğunluk <= 0.9975, Klorürler <= 0.044, Klorürler <= 0.043, Sitrik.asit <= 0.42: Şarap iyi

Kural24: Yoğunluk > 0.9959, Yoğunluk <= 0.9979, Yoğunluk <= 0.9975, Klorürler <= 0.044, Klorürler <= 0.043, Sitrik.asit > 0.42: Şarap Kötü

Kural25: Klorürler <= 0.043, Klorürler > 0.043: Şarap iyi

Kural26: Klorürler > 0.044, Uçucu.asitlik <= 0.28, Sitrik.asit <= 0.38, Klorürler <= 0.053, Uçucu.asitlik <= 0.22: Şarap iyi

Kural27: Uçucu.asitlik > 0.22, Uçucu.asitlik <= 0.245: Şarap Kötü

Kural28: Uçucu.asitlik > 0.22, Uçucu.asitlik > 0.245: Şarap iyi

Kural29: Klorürler <= 0.053, Klorürler > 0.053: Şarap Kötü

Kural30: Sitrik.asit <= 0.38, Sitrik.asit > 0.38: Şarap iyi

Kural31: Uçucu.asitlik <= 0.28, Uçucu.asitlik > 0.28: Şarap Kötü

Kural32: Yoğunluk > 0.9975, Klorürler <= 0.038: Şarap Kötü

Kural33: Yoğunluk > 0.9979, Sitrik.asit <= 0.37, Sitrik.asit <= 0.33: Şarap Kötü

Kural34: Yoğunluk > 0.9979, Sitrik.asit <= 0.37, Sitrik.asit > 0.33: Şarap iyi

Kural35: Yoğunluk > 0.9979, Sitrik.asit <= 0.37, Sitrik.asit > 0.37:Şarap Kötü

Tablo 3: C4.5 performans değerlendirme ölçütleri

Performans Değerlendirme Ölçütleri	C4.5
Doğruluk oranı	92.4686
Hata oranı	7.6318

Tablo 4: C4.5 kontenjans tablosu

C4.5	Gerçek			
Tahmin		Kötü	İyi	Kaliteli
	Kötü	88	10	1
	İyi	1	100	3
	Kaliteli	1	2	33

C4.5 karar ağacı algoritması kontenjans tablosu sonuçları incelendiğinde, İyi pozitif değerinin 88 çıktığı görülmektedir. Yani gerçekte Şarap Kalitesi Kötü olan model tahminde Şarap kalitesi 88 kişiyi doğru tahmin etmiştir.

Gerçekte Şarap Kalitesi, Kötü 1 pozitif değeri olan model, Şarap kalitesi iyidir şeklinde tahmin etmiştir. Yanlış pozitif değeri 1'dir.

Gerçekte Şarap Kalitesi, Kötü pozitif değeri 1 olan model, Şarap kalitesi Kaliteli şeklinde tahmin etmiştir. Yanlış pozitif değeri 1'dır.

Gerçekte Şarap Kalitesi, İyi pozitif değeri 10 olan model, Şarap Kalitesi kötü şeklinde tahmin etmiştir.Yanlış pozitif değeri 10'dur.

Gerçekte Şarap Kalitesi, İyi pozitif değeri 100 olan model, Şarap Kalitesi İyi şeklinde tahmin etmiştir. Doğru pozitif değeri 100'dür

Gerçekte Şarap Kalitesi, İyi pozitif değeri 2 olan model, Şarap Kalitesi Kaliteli şeklinde tahmin etmiştir. Yanlış pozitif değeri 2'dir.

Gerçekte Şarap Kalitesi, Kaliteli pozitif değeri 1 olan model, Şarap Kalitesi Kötüdür şeklinde tahmin etmiştir. Yanlış pozitif değeri 1'dir.

Gerçekte Şarap Kalitesi, Kaliteli pozitif değeri 3 olan model, Şarap Kalitesi İyi şeklinde tahmin etmiştir. Yanlış pozitif değeri 3'dür.

Gerçekte Şarap Kalitesi, Kaliteli pozitif değeri 33 olan model, Şarap Kalitesi Kaliteli şeklinde tahmin etmiştir. Doğru pozitif değeri 33'dür.

3.5 K-NN Doğruluk Oranı İçin C4.5 (Karar Ağacı) UYGULAMASI

K-NN algoritması doğruluk ve hata oranı belirlemek için K-NN'de oluşturduğumuz nümerik değerlerden ve hedef niteliği "Şarap Kalitesi Değeri" olarak belirlenen değişkenlerden oluşan bir alt küme belirlenir ve belirlenen bu alt küme ile C4.5 algoritmasında karar ağacı oluşturularak bu alt kümenin doğruluk ve hata oranı belirlenmiştir.

Nümerik değişkenlerden ve Hedef niteliği belirlenerek bu değişkenlerden oluşan bir alt küme daha oluşturulmuştur. Oluşturulan bu alt kümedeki değişkenler sırası ile şöyledir; "Uçucu asitlik"," Sitrik asit"," Klorürler"," Serbest kükürt dioksit"," Toplam kükürt dioksit", "Yoğunluk", "Sülfatlar" ve "Kalite".

Formülü uyguladığımızda 239 veri ve 8 değişkenden oluşan bir data.frame ulaşılır.

Hedef nitelik Şarap Kalitesi Değeri 4,5=Şarap Kötü , 6= Şarap İyi ve 7,8=Şarap Kaliteli şekline dönüştürülür.

Örneğin; Veri$Kalite <- revalue(veriler$Kalite, c("4"="Kötü,"5"="Kötü")) ' şeklindedir.

Şekil 21: K-NN ve C4.5 Summary

```
=== Summary ===

Correctly Classified Instances         221               92.4686 %
Kappa statistic                          0.8779
Mean absolute error                      0.078
Root mean squared error                  0.1975
Relative absolute error                 18.9649 %
Root relative squared error             43.5704 %
Total Number of Instances              239
```

Burada correctly classified instances doğru yerleşen tahmin sayısıdır. Bunun toplam 239 kişi içerisinden 221 kişi olduğu gözükmektedir ve %92.4 doğruluk oranına sahiptir.

Tablo 5: K-NN ve C4.5 Performans değerlendirme ölçütleri

Performans Değerlendirme Ölçütleri	**C4.5**
Doğruluk oranı	**%92.4686**
Hata oranı	**%7.532**

3.6 Naive – Bayes Sınıflandırıcı Algoritması

Naive Bayes sınıflandırması olasılık ilkelerine göre tanımlanmış bir dizi hesaplama ile, sisteme sunulan verilerin sınıfını yani kategorisini tespit etmeyi amaçlar. Naive 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 işletilir ve 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.

Analiz öncesi değişkenler faktör ve nümerik olarak tanımlanmış, nümerik veriler normalize edilip analize uygun hale getirilmiştir. Veri seti eğitim veri seti ve test veri seti olarak ayrılmıştır. Eğitim veri seti %60, test veri seti %40 olarak bölünmüştür. Eğitim ve test veri setine tahmininde kullanılacak nitelik ve hedef nitelik(Şarap Kalitesi) atanmıştır. Model tahmin edilmiş ve aşağıdaki koşullu olasılık değerleri bulunmuştur.

Call: naiveBayes.default(x = egitimNitelikleri, y = egitimHedefNitelik)

Şekil 22: Naive – Bayes Classifier

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = egitimNitelikleri, y = egitimHedefNitelik)

A-priori probabilities:
egitimHedefNitelik
     Kötü       İyi  Kaliteli
0.4137931 0.4344828 0.1517241

Conditional probabilities:
                 Uçucu.asitlik
egitimHedefNitelik      [,1]       [,2]
          Kötü     0.3110833 0.10959127
          İyi      0.2675397 0.09828314
          Kaliteli 0.3059091 0.15041717

                 Sitrik.asit
egitimHedefNitelik      [,1]      [,2]
          Kötü     0.3380000 0.1479899
          İyi      0.3612698 0.1115939
          Kaliteli 0.3390909 0.1104066

                 Klorürler
egitimHedefNitelik       [,1]        [,2]
          Kötü     0.04918333 0.010406145
          İyi      0.05269841 0.022996237
          Kaliteli 0.03940909 0.009550164

                 Serbest.kükürt.dioksit
egitimHedefNitelik     [,1]      [,2]
          Kötü     38.47500 19.039972
          İyi      39.11111 14.818375
          Kaliteli 34.86364  9.083189

                 Toplam.kükürt.dioksit
egitimHedefNitelik     [,1]     [,2]
          Kötü     150.7750 45.27547
          İyi      153.5238 43.95667
          Kaliteli 119.0455 37.40573

                 Yoğunluk
egitimHedefNitelik      [,1]        [,2]
          Kötü     1.1458300 1.162707400
          İyi      0.9951921 0.002426221
          Kaliteli 0.9915091 0.001835603

                 Sülfatlar
egitimHedefNitelik      [,1]       [,2]
          Kötü     0.4561667 0.10400606
          İyi      0.4811111 0.10780974
          Kaliteli 0.4745455 0.09414919
```

Tahmin edilen değerlerin ve gerçek değerlerin kıyaslanması için kontenjans tablosu elde edilmiştir.

Tablo 6: Naive – Bayes kontenjans tablosu

```
                 Gercek Siniflar
TahminiSiniflar Kötü İyi Kaliteli
        Kötü       0   0        0
        İyi       33  32        1
        Kaliteli   6   9       13
> |
```

Tablo 7: Naive bayes performans değerlendirme ölçütleri

Performans Değerlendirme Ölçütleri	Naive Bayes
Doğruluk oranı	**%23.4**
Hata oranı	**%76.5**

Naive Bayes algoritması kontenjans tablosu sonuçlarına göre gerçekte Şarap Kalitesi Kötü olan 0 veri, Tahminde de Şarap kalitesi Kötü olarak tahmin edilmiştir. Doğru pozitif değeri 0'dır.

Gerçekte Şarap Kalitesi Kötü olan 33 veri, Tahminde Şarap kalitesi iyi olarak tahmin edilmiştir. Yanlış pozitif değeri 33'dür.

Gerçekte Şarap Kalitesi Kötü olan 6 veri, Tahminde Şarap kalitesi Kaliteli olarak tahmin edilmiştir. Yanlış pozitif değeri 6'dır.

Gerçekte Şarap Kalitesi İyi olan 0 veri, Tahminde Şarap kalitesi Kötü olarak tahmin edilmiştir. Yanlış pozitif değeri 0'dır.

Gerçekte Şarap Kalitesi İyi olan 32 veri, Tahminde de Şarap kalitesi İyi olarak tahmin edilmiştir. Doğru pozitif değeri 32'dir.

Gerçekte Şarap Kalitesi İyi olan 9 veri, Tahminde Şarap kalitesi Kaliteli olarak tahmin edilmiştir. Yanlış pozitif değeri 9'dur.

Gerçekte Şarap Kalitesi Kaliteli olan 0 veri, Tahminde Kötü olarak tahmin edilmiştir. Yanlış pozitif değeri 0'dır.

Gerçekte Şarap Kalitesi Kaliteli olan 1 veri, Tahminde iyi olarak tahmin edilmiştir. Yanlış pozitif değeri 1'dir.

Gerçekte Şarap Kalitesi Kaliteli olan 13 veri, Tahminde Kaliteli olarak tahmin edilmiştir. Doğru pozitif değeri 13'dür.

3.7 Genel Değerlendirme ve Model Seçimi

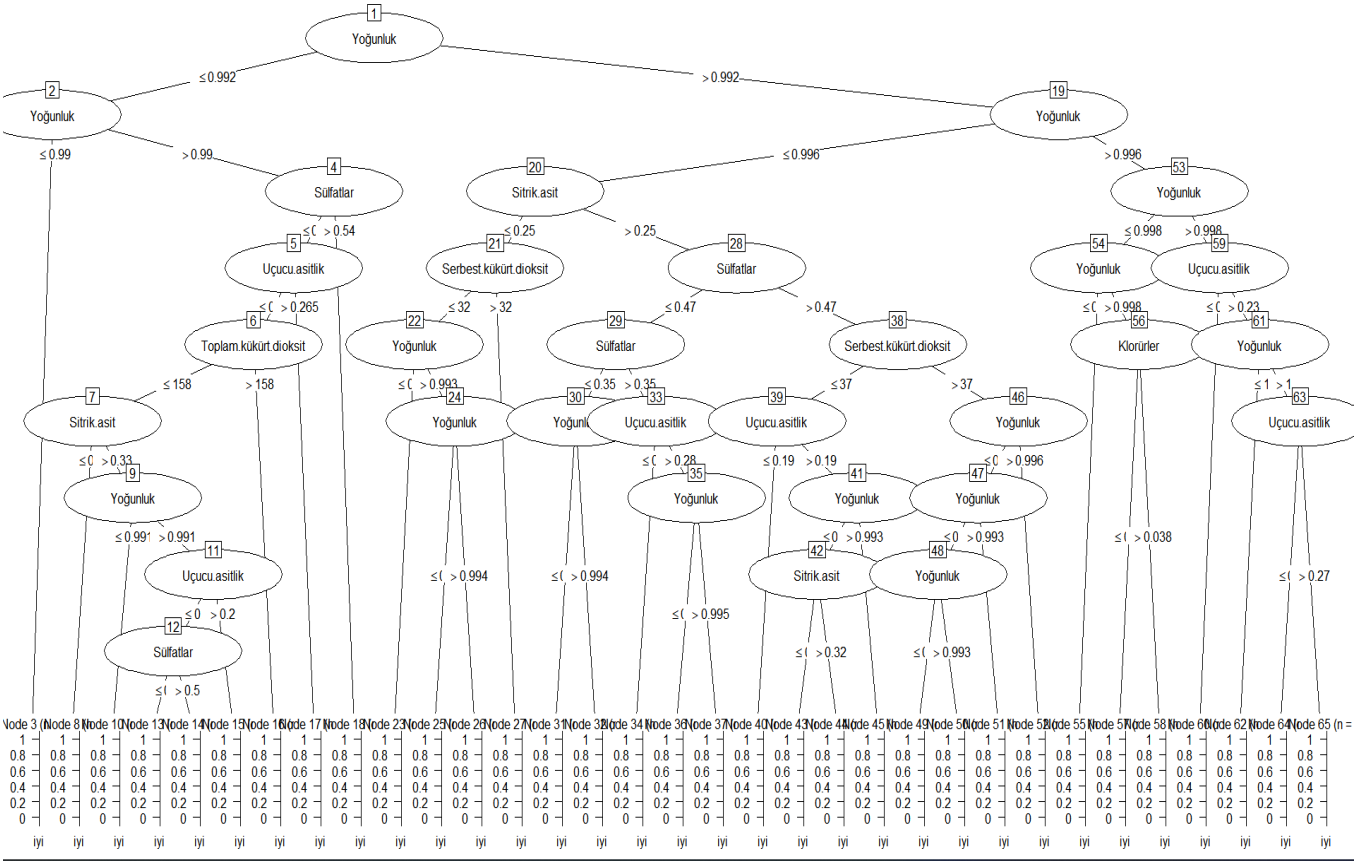
Şarap Kalitesi değerlendirme için Sırasıyla KNN, C4.5 Karar Ağacı ve Naive Bayes Algoritmaları kullanılmış ve bu algoritmaların performans değerlendirme ölçütleri kıyaslanmıştır.

Tablo 7 : Genel değerlendirme ve model seçimi

	Doğruluk	Hata
KNN	%92.4686	7.532
C4.5 Karar Ağacı	%92.4686	7.532
Naive Bayes Algoritması	%20.2	79.8

Belirlenen performans değerlendirme oranlarına göre KNN, C4.5 karar ağacı algoritması aynı derecede çıkmıştır. Naive Bayes Algoritması ise onların tam tersi çıkmıştır. Doğruluk ve hata oranı baz alınırsa en iyi performans veren algoritma KNN ve C 4.5 algoritmasıdır denilebilir.

Şekil 23: C4.5 Karar Ağacı Şekli



SONUÇ

Kalitesiz alkol insan vücudunda yaralara, kör olmasına ve en önemlisi hayatına sebep olabilir. Kalitesiz içki Türkiye'de 2021'in, Aralık ayında 22 ilde 89 kişinin hayatını kaybetmesine sebep olmuştur. Kalitesiz alkol yaşam kalitesini azaltmakta ve sağlık harcamalarını arttırmakta olan bir hastalık olmaya başlamıştır.

Kalitesiz alkolün dünya üzerinde yaygın olarak görülmesi, tedavi edilmediğinde veya tedavide geç kalınması durumunda organlara zarar vermesi, önemli bir tehdittir. Konunun önemi sebebiyle Final proje ödevimde, Kalitesiz Şarabı etkileyen faktörlerin belirlenmesi hedeflenmiştir.

Bu projede veri madenciliğinden yararlanarak Şarap Kalitesi üzerine çalışma yapılmıştır 3 farklı yöntemden yararlanılmıştır.

Bunlar sırasıyla şöyledir; K-NN, C 4.5 Karar Ağacı Algoritması ve Naive Bayes'tir.

Projenin birinci bölümünde veri madenciliği kavramı ve veri madenciliğinin uygulandığı alanlar ele alınmıştır.

İkinci bölümde Veri madenciliğinin aşamalarına yer verilmiştir.

Projenin üçüncü bölümü uygulama bölümüdür. Veri madenciliği sürecine sadık kalınarak, uygulama aşamaları anlatılmış ve uygulamada kullanılan tekniklere yer verilmiştir. K-nn, C4.5 ve Naive-Bayes Algoritması kullanılmıştır.

K-nn algoritmasın da gerçek veriler ile Şarap kalitesi yapılmaya çalışıldığından Şaraplar 4 ile 8 arasında puanlandırılmıştır. 4-5 arası Şarap kalitesi "Kötü", 6 Şarap kalitesi "İyi" ve 7-8 arası Şarap Kalitesi "Kaliteli" olarak belirlenmiştir. Veri setinden rast gele çağırdığım 30 kişiyi En yakın komşu algoritmasında ağaçlandırma yoluyla Türlerine ayırarak kendi içlerinde gruplanma yaptım ve 2 farklı Grafik oluşturdum. Veri setinde bulunan 239 kişi içerisinden rast gele 30 kişi çağırılmıştır. Buna göre doğrulur Oranı **%92.4686 ve hata oranı %7.532'dir.**

C4.5 karar ağacı algoritması karışık matrisi sonuçları incelendiğinde. "Doğru pozitif" değeri 88'dir, "Yanlış pozitif" 1'dir, "Yanlış pozitif "değeri 1'dir, "Yanlış pozitif" değeri 1'dir, "Yanlış pozitif" değeri 10'dur, "Doğru pozitif" değeri 100'dür, "Yanlış pozitif" değeri 2'dir. "Yanlış pozitif" değeri 1'dir, "Yanlış pozitif" değeri 3'dür. "Doğru pozitif" değeri 33'dür. Tahmin edilmiştir.

Modelin performans ölçümlerine bakıldığında doğruluk oranı **92.4686 ve hata oranı %7.532'dir.**

C4.5 karar ağacı algoritması ile ortaya çıkan 35 tane kural vardır. Bu kurallar içerisindeki belirleyici değişkenler Serbest kükürt dioksit, Yoğunluk, Sitrik asit ve Uçucu asit değişkenleridir.

Naive Bayes algoritması sonuçları incelendiğinde gerçekte Şarap Kalitesi Kötü olan modelin 0 Tahminde de Şarap kalitesi Kötü olarak tahmin edilmiştir. Doğru pozitif değeri 0'dır.

Gerçekte Şarap Kalitesi Kötü olan 33 veri, Tahminde Şarap kalitesi iyi olarak tahmin edilmiştir. Yanlış pozitif değeri 33'dür.

Gerçekte Şarap Kalitesi Kötü olan 6 veri, Tahminde Şarap kalitesi Kaliteli olarak tahmin edilmiştir. Yanlış pozitif değeri 6'dır.

Gerçekte Şarap Kalitesi İyi olan 0 veri, Tahminde Şarap kalitesi Kötü olarak tahmin edilmiştir. Yanlış pozitif değeri 0'dır.

Gerçekte Şarap Kalitesi İyi olan 32 veri, Tahminde de Şarap kalitesi İyi olarak tahmin edilmiştir. Doğru pozitif değeri 32'dir.

Gerçekte Şarap Kalitesi İyi olan 9 veri, Tahminde Şarap kalitesi Kaliteli olarak tahmin edilmiştir. Yanlış pozitif değeri 9'dur.

Gerçekte Şarap Kalitesi Kaliteli olan 0 veri, Tahminde Kötü olarak tahmin edilmiştir. Yanlış pozitif değeri 0'dır.

Gerçekte Şarap Kalitesi Kaliteli olan 1 veri, Tahminde iyi olarak tahmin edilmiştir. Yanlış pozitif değeri 1'dir.

Gerçekte Şarap Kalitesi Kaliteli olan 13 veri, Tahminde Kaliteli olarak tahmin edilmiştir. Doğru pozitif değeri 13'dür.

Naive Bayes algoritması performans ölçümlerine bakıldığında doğruluk oranı **%20.2** ve hata oranı **79.8**'tir.

Tüm modeller birlikte değerlendirildiğinde performans ölçüm modelleri değerlendirme ölçütlerine göre en yüksek doğruluk ve en düşük hatayı veren C4.5 ve K-NN algoritması en uygun modellerdir denilebilir.

KAYNAKÇA

https://dergipark.org.tr/en/download/article-file/1407214

https://dergipark.org.tr/en/download/article-file/991964

http://kergun.baun.edu.tr/veri_madenciligi_hafta5.pdf

https://kodedu.com/2014/05/naive-bayes-siniflandirma-algoritmasi/

https://www.acarindex.com/dosyalar/makale/acarindex-1423940013.pdf

https://www.gtech.com.tr/veri-madenciligi-nedir-ve-nasil-yapilir/

http://mgocenoglu.blogspot.com/2014/06/veri-madenciligi-asamalar.html

https://medium.com/@ipekkrdmn/veri%CC%87-madenci%CC%87li%CC%87%C4%9Fi%CC%87-nedi%CC%87r-7a8d936eff95

https://web.karabuk.edu.tr/emelkocak/indir/MTM326/veri_madencili%C4%9Fi.pdf

file:///C:/Users/user/Desktop/Nur%20Kuban%20Torun%20Doktora%20Tez.pdf

EKLER

Ek 1: Veri Önişleme İçin Kullanılan R kodlar

#Kullanılan veri seti dosyadan seçilir.

> veriler =read.table (file.choose(),header=T,sep=";")

#Veri yapısı incelenir.

> str(veriler)

'data.frame': 239 obs. of 8 variables:

$ Uçucu.asitlik : num 0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...

$ Sitrik.asit : num 0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...

$ Klorürler : num 0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...

$ Serbest.kükürt.dioksit : num 45 14 30 47 47 30 30 45 14 28 ...

$ Toplam.kükürt.dioksit : num 170 132 97 186 186 97 136 170 132 129 ...

$ Yoğunluk : num 1.001 0.994 0.995 0.996 0.996 ...

$ Sülfatlar : num 0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...

$ Kalite : Factor w/ 5 levels "4","5","6","7",..: 3 3 3 3 3 3 3 3 3 3 ...

Hedef nitelik, Şarap Kalitesi değişkeninin değerleri 4,5= kötü, 6= iyi, 7,8= Kaliteli şekline dönüştürülür.

```
> install.packages("plyr")

> library(plyr)

veriler$Kalite <- revalue(veriler$Kalite, c("4"="Kötü","5"="Kötü"))

veriler$Kalite <- revalue(veriler$Kalite, c("6"="İyi"))

veriler$Kalite <- revalue(veriler$Kalite, c("7"="Kaliteli","8"="Kaliteli"))


#Veri setinin özetine bakılır

 > summary(veriler)


#Nümerik değişkenlerin grafikleri çizilir.

hist(veriler$Klorürler, col="red", main = "Klorürler Histogram Grafiği")

hist(veriler$Serbest.kükürt.dioksit, col="red", main = "Serbest kükürt dioksit Histogram
Grafiği")

hist(veriler$Toplam.kükürt.dioksit, col="red", main = "Toplam kükürt dioksit Histogram
Grafiği")

hist(veriler$Yoğunluk, col="red", main = "Yoğunluk Histogram Grafiği")

hist(veriler$Sülfatlar, col="red", main = "Sülfatlar Histogram Grafiği")

hist(veriler$Uçucu.asitlik, col="red", main = "Uçucu asitlik Histogram Grafiği")

hist(veriler$Sitrik.asit, col="red", main = "Sitrik Asit Grafiği")


#kutu grafikleri çizimi

boxplot(veriler$Klorürler, col="orange", main="Klorürler Kutu Grafiği")

boxplot(veriler$Serbest.kükürt.dioksit, col="orange", main="Serbest kükürt dioksit Kutu
Grafiği")

boxplot(veriler$Toplam.kükürt.dioksit, col="orange", main="Toplam kükürt dioksit Kutu
Grafiği")

boxplot(veriler$Yoğunluk, col="orange", main="Yoğunluk Kutu Grafiği")

boxplot(veriler$Sülfatlar, col="orange", main="Sülfatlar Kutu Grafiği")

boxplot(veriler$Sitrik.asit, col="orange", main="Sitrik asit Grafiği")

boxplot(veriler$Uçucu.asitlik, col="orange", main="Uçucu asitlik Kutu Grafiği")
```

#serpilme diyagramı çizimi

pairs(~ Uçucu.asitlik+ Sitrik.asit + Klorürler + Serbest.kükürt.dioksit +Toplam.kükürt.dioksit + Yoğunluk + Sülfatlar , data= veriler, col=" dark green", main= "Serpilme Diyagramları")

Ek2: KNN Algortiması Uygulaması Kodları

#kümeleme analizi knn

- veriler= read.table(file.choose(), header = T, sep = ";")
- library(cluster)

- library(plyr)
- data(veriler)
- View(veriler)
- summary(veriler)
- str(veriler)
- attributes(veriler)

- veriler$Uçucu.asitlik <- as.numeric(veriler$Uçucu.asitlik)
- veriler$Sitrik.asit <- as.numeric(veriler$Sitrik.asit)
- veriler$Klorürler <- as.numeric(veriler$Klorürler)
- veriler$Serbest.kükürt.dioksit <- as.numeric(veriler$Serbest.kükürt.dioksit)
- veriler$Toplam.kükürt.dioksit <- as.numeric(veriler$Toplam.kükürt.dioksit)
- veriler$Yoğunluk <- as.numeric(veriler$Yoğunluk)
- veriler$Sülfatlar <- as.numeric(veriler$Sülfatlar)
- veriler$Kalite <- as.character(veriler$Kalite)

- veriler$Kalite <- revalue(veriler$Kalite, c("4"="Kötü","5"="Kötü"))
- veriler$Kalite <- revalue(veriler$Kalite, c("6"="İyi"))

- veriler$Kalite <- revalue(veriler$Kalite, c("7"="Kaliteli","8"="Kaliteli"))

#rastalantisal olarak 239 veriden 30 tanesini çekelim

- library("caret")
- set.seed(1234)

#sample fonksiyonu ile tesadüfi sayıyı elde edeceğiz

- ind <- sample(1:239:30)
- veri <- veriler[ind,]

en yakın komsu algoritması

- modelo <- agnes(veri,metric = "euclidien", method = "single") #"oklit uzaklığına göre”

#Grafikte Gösterelim

pltree(modelo, main="en yakin komsu algoritması ile kumeleme")

#Bu görselde sayılar ile gösterim var, sinif etiketi seklinde göstermek istersek

- pltree(model, main="en yakın komsu algoritması ile kumeleme",labels=veri$Kalite)

#sonucu banner grafik seklinde gösterelim

- bannerplot(agnes(veri), main ="bannerplot Grafiği", labels = veri$Kalite)

Ek 3:C4.5 Karar Ağacı Algoritması Kodları

#C4.5 Karar ağacı Algoritması

- veriler= read.table(file.choose(), header = T, sep = ";")
- library(rJava)
- library(RWeka)
- head(veriler)
- data(veriler)
- View(veriler)
- str(veriler)
- summary(veriler)
- install.packages("Plyr")
- library(plyr)

- data(veriler)

- veriler$Kalite <- as.factor(veriler$Kalite)

- veriler$Kalite <- revalue(veriler$Kalite, c("4"="Kötü","5"="Kötü"))
- veriler$Kalite <- revalue(veriler$Kalite, c("6"="iyi"))
- veriler$Kalite <- revalue(veriler$Kalite, c("7"="Kaliteli","8"="Kaliteli"))

- veriler$Kalite <- as.factor(veriler$Kalite)

- model <- J48(Kalite~.,data = veriler)

- View(model)
- print(model)
- summary(model)
- plot(model)
- summary(veriler)

Ek 4: Naive – Bayes Algortiması İçin Kodlar

#Naive Bayes Algoritması

#Önce veri seti çağırıldı

- veriler =read.table (file.choose(),header=T,sep=";")

veri seti incelenir, nümerik ve kategorik veriler tananımlanır

- library("plyr")
- veriler$Kalite <- as.factor(veriler$Kalite)
- veriler$Kalite <-revalue(veriler$Kalite, c("4"="Kötü","5"="Kötü"))
- veriler$Kalite <-revalue(veriler$Kalite, c("6"="İyi"))
- veriler$Kalite <-revalue(veriler$Kalite, c("7"="Kaliteli","8"="Kaliteli"))

#veri seti eğitim ve test veri seti olarak ayrılır.

#veri seti egitim ve test seti olarak ikiye ayrilacak

- library(caret)

- set.seed

- verisetibolme <- createDataPartition(y=veriler$Kalite, p=0.6,list=FALSE)

- egitim <- veriler[verisetibolme,]

- test <- veriler[-verisetibolme,]

#Eğitim ve test veri setine tahmininde kullanılacak nitelik ve hedef nitelik(diyabetik

#polinöropati) atanır. Diyabetik polinöropati 8. Sütunda olduğu için 8 kullanıldı.

- testNitelikleri <- test[,-8]
- testHedefNitelik <- test[[8]]
- egitimNitelikleri <- egitim [,-8]
- egitimHedefNitelik <- egitim [[8]]

- library(e1071)

- naiveBayes_modeli_kuruldu <- naiveBayes(egitimNitelikleri, egitimHedefNitelik)

- naiveBayes_modeli_kuruldu

#modelin tahminleri bulunur

- (tahminiSiniflar <- predict(naiveBayes_modeli_kuruldu, testNitelikleri))

#gercek siniflar ile tahmini siniflarin kıyasi

- (karisiklikmatrisi <- table(tahminiSiniflar, testHedefNitelik, dnn =c ("TahminiSiniflar", "Gercek Siniflar")))

- (TP <- karisiklikmatrisi [1])

- (FP <- karisiklikmatrisi [4])

- (FN <- karisiklikmatrisi [7])

- (TN <- karisiklikmatrisi [2])
- (TP <- karisiklikmatrisi [5])
- (FP <- karisiklikmatrisi [8])
- (FN <- karisiklikmatrisi [3])
- (TN <- karisiklikmatrisi [6])
- (TP <- karisiklikmatrisi [9])
- paste0("Dogruluk = ",(Dogruluk <- (TP+TN)/sum(karisiklikmatrisi)))
- paste0("Hata = ",(Hata <- 1-Dogruluk))

Ek 5: KNN KARAR AĞACINA UYARLANMASI

knn karar ağacına uyarlanmıştır.Sadece nümerik değerlerden oluşmaktadır.

- veriler= read.table(file.choose(), header = T, sep = ";")
- install.packages("caret")
- library(caret)
- install.packages("cluster")
- library("cluster")
- library("plyr")
- library("RWeka")
- library("rJava")
- View(veriler)
- summary(veriler)

- str(veriler)

- attributes(veriler)

#VERİLER nümerik ve faktör olarak tanımlanır

- veriler$Uçucu.asitlik <- as.numeric(veriler$Uçucu.asitlik)
- veriler$Sitrik.asit <- as.numeric(veriler$Sitrik.asit)
- veriler$Klorürler <- as.numeric(veriler$Klorürler)
- veriler$Serbest.kükürt.dioksit <- as.numeric(veriler$Serbest.kükürt.dioksit)
- veriler$Toplam.kükürt.dioksit <- as.numeric(veriler$Toplam.kükürt.dioksit)
- veriler$Yoğunluk <- as.numeric(veriler$Yoğunluk)
- veriler$Sülfatlar <- as.numeric(veriler$Sülfatlar)
- veriler$Kalite <- as.character(veriler$Kalite)

- veriler$Kalite <- revalue(veriler$Kalite, c("4"="Kötü","5"="Kötü"))
- veriler$Kalite <- revalue(veriler$Kalite, c("6"="İyi"))
- veriler$Kalite <- revalue(veriler$Kalite, c("7"="Kaliteli","8"="Kaliteli"))

#sadece nümerik değerlerden oluşan alt küme oluşturuldu ve nümerik değerlere karşılık gelen

#verilerin değerleri sayısal olarak girildi.

- n_veriler <- veriler [c(1,2,3,4,5,6,7,8)]

#rastgele veri seçimi için set.seed kullanılır.

- set.seed(1234)
- ind <- sample(1:239,239)
- veriler <- n_veriler[ind,]
- veriler$Kalite <- as.factor(veriler$Kalite)
- deneme <- J48(Kalite~.,data = veriler)#kurallari görelim
- print(deneme)
- summary(deneme)

#grafigini çizelim

- plot(deneme)