T.C.

FATİH SULTAN MEHMET VAKIF ÜNİVERSİTESİ

Mühendislik Fakültesi

Bilgisayar Mühendisliği Bölümü

Lisans Bitirme Projesi

DOĞRUSAL, DOĞRUSAL OLMAYAN REGRESYON VE SINIFLANDIRMA YARDIMIYLA ORMAN YANGINI TAHMİNİ

YİĞİT MERT DÖNMEZ

1421221016

Bitirme Projesi Danışmanı: DR. ÖĞR. ÜYESİ ZEYNEP GÜNDOĞAR

İstanbul, [Mayıs 2019]

T.C.

FATİH SULTAN MEHMET VAKIF ÜNİVERSİTESİ

Mühendislik Fakültesi

Bilgisayar Mühendisliği Bölümü

Lisans Bitirme Projesi

DOĞRUSAL, DOĞRUSAL OLMAYAN REGRESYON VE SINIFLANDIRMA YARDIMIYLA ORMAN YANGINI TAHMİNİ

YİĞİT MERT DÖNMEZ

1421221016

Bitirme Projesi Danışmanı DR. ÖĞR. ÜYESİ ZEYNEP GÜNDOĞAR

Jüri Üyeleri:	İmza:
Prof. Dr. Ali Yılmaz ÇAMURCU	
Dr. Öğr. Üyesi Zeynep GÜNDOĞAR	
Dr. Öğr. Üyesi Berna KİRAZ	

İstanbul, [Mayıs 2019]

ONAY SAYFASI

Danışman : Prof. Dr. Ali Yılmaz ÇAMURCU **..…….....**

Juri Üyeleri:

Prof. Dr. Ali Yılmaz ÇAMURCU **.....**

Dr. Öğr. Üyesi Zeynep GÜNDOĞAR **.....**

Dr. Öğr. Üyesi Berna KİRAZ **.....**

ÖZET

Bu projede büyük verilerde doğrusal, doğrusal olmayan regresyon ve sınflandırma yardımıyla orman yangını tahmini konusu ele alınmıştır. Projenin ana amacı ülkelerde bulunan orman yangınlarını engelleyebilmektir. Bu projede kullanılan veri seti Portekiz sınırları içerisinde hazırlanmıştır. İlk olarak veri setine temel bileşenler analizi uygulanıp boyut indirgenmiştir. Temel bileşenler analizinden sonra regresyon teknikleri olan doğrusal regresyon ve doğrusal olmayan regresyon analizi uygulanmıştır. Son olarak veri seti üzerine sınıflandırma teknikleri uygulanıp sonuçlar karşılaştırılmıştır.

ABSTRACT

In this project, the estimation of forest fire with the help of linear regression, nonlinear regression and classification in large data were taken into consideration. The main purpose of the project is to prevent forest fires in countries. The data set used in this project was prepared within the borders of Portugal. Firstly, the principle component analysis was applied to the data set and the dimension was reduced. After the principle component analysis, linear regression and nonlinear regression analysis techniques were applied. Finally, classification techniques were applied on the data set and the results were compared.

ÖNSÖZ

Bu bitirme projesinde Portekiz'de toplanmış orman yangınları verisinin, regresyon analizi teknikleri, Sınıflandırma teknikleri ile büyük veri teknolojisi kullanarak modellenmesi ve gelecek analizinin yapılması amaçlanmıştır.

Günümüzde giderek çoğalmakta olan orman yangınları bu projenin gerçekleştirilmesinin önemini daha da arttırmaktadır. Proje gerçekleştirildikten sonra diğer ülkelerde de veri kullanımı sağlanabildiği sürece kullanılabilir.

Bu projede Fatih Sultan Mehmet Vakıf Üniversitesi, proje danışmanlarım Dr. Öğr. Üyesi Süha TUNA ve Dr. Öğr. Üyesi Zeynep GÜNDOĞAR'a teşekkür ederim.

05.2019 YİĞİT MERT DÖNMEZ

İÇİNDEKİLER

Sayfa

Şekil Listesi

Sayfa

Tablo Listesi

Sayfa

1. GİRİŞ

1.1. Orman Yangınları ve Önemi

Orman yangınları günümüzde büyük sorunlar ortaya çıkarabiliyor. Özellikle yaz aylarında çoğu Akdeniz ülkesinde orman yangınları daha fazla oluşabiliyor. Orman yangınlarının sadece ısınan havanın ve nemin azalmasıyla oluştuğu gerçeği kabul edilebilir bir gerçek değildir. Orman yangınlarıyla alakalı istatistiklere bakıldığında yangın sebeplerinin iklimin etkisinden daha çok insan kaynaklı çıktığı görülmektedir. Türkiye'de yapılan bir araştırmada yangınların yaklaşık %94'ü insan kaynaklı oluşmaktadır. Bunların yaklaşık %13'ü tarla, arsa açmak niyetiyle çıkarılan yangınlardır. Yüzde 47'si ihmal, dikkatsizlik ve kazadan kaynaklanıyor. Geriye kalan %34'ü ise insan kaynaklı olmasına karşın neden çıktığı bilgisi bilinmemektedir. İhmal ve dikkatsizlik sonucu çıkan yangınlar arasında anız yakma, çoban ateşi ve sigara önemli yer tutarken son yıllarda, enerji nakil hatlarından kaynaklanan yangınlar da çok önemli artışlar görülüyor. İlerleyen dönemlerde iklimsel olarak ısının daha da artacağı ve yağışların, nemin daha da azalacağı tahmin ediliyor. Bu etkilerde orman yangınlarının artacağı yönünde bilgi vermektedir [1].

Orman yangınlarını azaltıcı etki olarak insanları bilinçlendirmek, doğanın insan üzerindeki etkisini insanlara anlatmak, insanların sebep olduğu orman yangınlarının önüne geçmeyi sağlayabilmektedir. Bu projede regresyon analizi ile orman yangınları verisi kullanılarak gelecek analizi yapılması amaçlanmıştır. Bu analiz sonucunda orman yangınlarında azaltıcı etki sağlanması istenmektedir [2].

1.2. Temel Bileşenler Analizi ile Veri İndirgeme

Temel bileşenler analizi çok değişkenli analizlerin temelini oluşturmaktadır. Temel bileşenler analizi, çok sayıda bağımlı değişkenin, olabildiğince daha az sayıda doğrusal bileşenlerine indirgenmesinde kolaylıklar sağlayan bir analiz tekniğidir.

Bu analizin en önemli amacı, korelasyon matrisinin, olabildiğince az bilgi kaybı vererek indirgenmesini sağlamak, daha az sayıda birbirinden bağımsız doğrusal bileşenleri olan, yeni değişkenlerin belirlenmesini sağlamaktır.

Temel bileşenler analizinde, orijinal değişkenlerden elde edilecek olan sonuçlardan çok az bir bilgi kaybı göz ardı edilebilecek şekilde indirgeme yapılırken, temel bileşenlerin sayısı orijinal değişken sayısından az olması temel bileşenler analizinin amacıdır. Dolayısıyla temel bileşenler analizi, daha az sayıda yeni değişken üretirken bilgi kaybını en az seviyede tutmaya çalışan bir analiz tekniğidir [3].

Temel bileşenler analizi, değişkenler arasındaki bağımlılık yapısını ortadan kaldırmak, boyut indirgemek ve çeşitli ölçütler bakımından sıralamalar yapmakta kullanılmaktadır.

Temel bileşenler analizi, bir analiz tekniği olduğu gibi, başka analiz teknikleri için de veri hazırlama tekniği olarak kullanılmaktadır. Temel bileşenler analizi, çok değişkenli analiz teknikleri olan Faktör Analizi, Ayırma Analizi, Lojistik Regresyon Analizi ve Çok Değişkenli Regresyon Analizi için de ilk adımı oluşturmaktadır. Bu proje için ise temel bileşenler analizi regresyon analizlerine veri hazırlama tekniği için kullanılacaktır [3].

Bu çalışmada yapılan uygulama örneğinde Portekiz'de daha önce çıkmış orman yangınlarının gelecekte Portekiz'in hangi bölgelerinde çıkabileceği bilgisinin doğrusal ve doğrusal olmayan regresyon analizi kullanarak tespit edilmesi amaçlanmıştır.

1.3. Doğrusal ve Doğrusal Olmayan Regresyon ile Veri Modelleme

Regresyon analizi, aralarında sebep-sonuç ilişkisi bulunan iki veya daha fazla değişken arasındaki ilişkiyi belirlemek ve bu ilişkiyi kullanarak o konu ile ilgili tahminler (estimation) ya da kestirimler (prediction) yapabilmek amacıyla yapılır. Regresyon analizi değişkenler arasındaki ilişki doğrusal ise, doğrusal (lineer) regresyon analizi, değil ise doğrusal-olmayan (nonlineer) regresyon analizi olarak isimlendirilir[4].

Doğrusal regresyon analizi, tek değişkenli doğrusal regresyon analizi ve çok değişkenli doğrusal regresyon analizi olarak ayrılmaktadır. Tek değişkenli regresyon analizi, bir bağımlı değişken ve bir bağımsız değişken arasındaki ilişkiyi inceleyen analiz tekniğidir.

Bu analizle bağımlı ve bağımsız değişkenler arsındaki doğrusal (lineer) ilişkiyi temsil eden bir doğru denklemi formüle edilmektedir. Çok değişkenli doğrusal regresyon analizinde ise bir bağımlı değişken ve birden fazla bağımsız değişkenin yer aldığı regresyon modeli bulunmaktadır. Çok değişkenli regresyon analizinde bağımsız değişkenler eş zamanlı olarak (aynı anda) bağımlı değişkendeki değişimi açıklamaya çalışmaktadır.

Doğrusal olmayan regresyon, gözlemsel verilerin model parametrelerinin doğrusal olmayan bir kombinasyonu olan ve bir veya daha fazla bağımsız değişkene bağlı olan bir fonksiyonla modellendiği bir regresyon analizi şeklidir . Veriler ardışık yaklaşımlar metodu ile donatılmıştır.

1.4. Sınıflandırma

Sınıflandırma, bir veri kümesinde bulunan değerlerin hangi sınıfa ait olduğunun bulunmasında kullanılmaktadır. Sınıflandırma algoritmalarında sistem veri setini eğitim verisi üzerinde işlem yaparak öğrenir. Daha sonra sistemin hiç bilmediği bir test verisi üzerinde sınıfı bilinmeyen değerlerin sınıfının bulunmasında kullanılır. Örnek olarak bu projede çeşitli özellikler aracılığı ile belirtilen yerlerde orman yangını olup olmadığı test edilmiştir.

2. TEMEL BİLEŞENLER ANALİZİ İLE VERİ İNDİRGEME

2.1. Çok Değişkenli İstatistiksel Analiz Tanımı

İstatistik biliminde kullanılan tek değişkenli istatistiksel analiz yöntemlerinin çeşitli durumlarda yeterli olmadığı görülerek çok değişkenli istatistiksel analiz yöntemleri kullanılmaya başlanmıştır. Çok değişkenli istatistiksel analiz yöntemleri kullanılarak değişkenler arası bağımlılığı ortadan kaldırmak, sınıflandırma yapmak ve boyut indirgemek hedeflenmektedir.

Herhangi bir durumun tanımlanması, sadece bir tane değişkenin etkisi ile değil çok daha fazla bağımlı veya bağımsız değişkenin etkisiyle karmaşık bir yapı göstermektedir. Bu yüzden bir tane değişkene göre değil birden fazla değişkene göre tanımlama yapılması gerekmektedir. Bu şekilde bir tanımlama yapılmadığı durumda değişkenler arası etkiler dikkate alınmamış olur.

2.1.1. Çok değişkenli istatistiksel analizin kullanımı ve önemi

İstatistiksel analizde kullanılan önemli yöntemlerden biri çok değişkenli istatistiksel analiz yöntemidir. Çok değişkenli istatistiksel analizde birbiriyle ilişkili çok sayıda değişken bulunmaktadır. Kullanılan tekniklerle, sistem basitleştirilerek herhangi bir konuda kesin sonuç için gerekli bilgi çıkartılır [3].

Çok değişkenli istatistiksel yöntemler, incelenen değişkenleri etkileyen faktörleri ve bu faktörlerin birbiriyle olan ilişkilerini ortaya çıkararak çok boyutlu sistemi orijinal değişkenlerin doğrusal bileşimi olan az sayıda bileşen ile özetlemektedir. Bu nedenle çok değişkenli yöntemler tek değişkenli yöntemlere göre daha karmaşık olabilmektedir.

Çok değişkenli analiz birden fazla özellikle ilgilendiğinden uygulamalarda değişik amaçlarda kullanılmaktadır. Bu amaçlar aşağıdaki gibidir.

1- **Basitleştirme ve Boyut İndirgeme:** Bir sistemde bulunan çok sayıda değişken, daha az sayıda değişkenle ifade edilir. Böylece basitlik sağlanmış olur.

2- **Birimlerin sınıflandırılması:** Verilerin değişik sınıflar oluşturup oluşturmadıkları gözlemlenir. Diğer bir yöntem ise birimlerin önceden tanımlanmış sınıflardan hangilerine ait olduklarının belirlenmesidir.

3- **Bağımlılık Yapısının İncelenmesi:** Değişkenlerin kovaryans ve korelâsyonlarından yararlanılarak bağımlılığın kaynakları ve sonuçları açıklanır. Örneğin; değişkenlerden bir ya da daha fazlasının bağımlı, ötekilerin bağımsız olduğu regresyon analizinin amacı, değişkenler arasındaki bağımlılık yapısını ortaya çıkarmaktır.

4- **Ölçekleme:** Ölçekleme çok sayıda değişkenden yararlanarak birimlerin daha az boyutla gösterilmesidir. Böylece, grafik gösterimlerinden de yararlanılarak birimlerin karşılaştırılması, kolaylıkla yapılabilmektedir.

2.1.2. Çok değişkenli istatistiksel analiz tekniklerinde temel bileşenler analizinin yeri

Çok değişkenli herhangi bir veri setinde p adet değişken arasındaki ilişki, eğer değişkenler arası bağımlı bağımsız değişken olarak ayrılabiliyorsa regresyon teknikleri ile incelenir. Bu ayrımın yapılamaması ve değişkenler arası birlikte değişimin yorumlanması gereksinimi çeşitli boyut indirgeme yöntemlerinin kullanılmasına neden olmuştur ki bu yöntemlerden biri de temel bileşenler analizidir.

Temel bileşenler analizinde n tane nesneye ilişkin p tane değişken incelenmektedir. Bu değişkenlerden birçoğunun birbirleriyle bağımlı ve nesne sayısının çok fazla olması analiz yaparken sorun yaratmaktadır. Örneğin bir kıyafetin özellikleri değişkenleri ifade ediyor olsun. Bu durumda bedeni, boyu, rengi, modeli, deseni vs. çok sayıda değişken bulunmaktadır. Bu değişkenlerin bazıları birbiriyle ilişkilidir. Fakat bu durum değişkenlerin bağımsızlığı kuralını çiğnemiş olur. Aynı zamanda, çok sayıda değişken ile işlem yapmak işlem yükünü arttıracağı ve elde edilecek sonuçların yorumunda bazı güçlüklere neden olacağı için tercih edilmek istenen bir durum değildir. Çok sayıda değişkene ait analiz sonuçlarının yorumlanması ve özetlenmesi günümüzde bile gerçekten zor olabilmektedir. Böyle durumlarda başvurulan tekniklerden en önemlisi Temel Bileşenler Analizi (Principal Component

Analysis)'dir. Temel bileşenler analizi değişkenler arası bağımlılık yapısının yok edilmesi ve/veya boyut indirgeme amacıyla kullanılmaktadır. Bunun yanı sıra başka analizler için veri hazırlama tekniği olarak da kullanılabilmektedir [3].

2.2. Temel Bileşenler Analizine Giriş

Temel bileşenler analizi, orijinal x değişkenlerinden oluşan veri setini, daha az sayıda ve bu değişkenlerin doğrusal bileşenleri olan yeni değişkenlerle ifade etmemizi sağlar. Aralarında korelâsyon bulunan x sayıda değişkeni, aralarında korelâsyon bulunmayan ve sayıca orijinal değişken sayısından daha az sayıda ($y<x$) orijinal değişkenlerin doğrusal bileşenleri olan y tane değişkenle ifade etme yöntemine temel bileşenler analizi denir.

Temel bileşenler analizinde bağımlı değişken bulunmamaktadır. Temel bileşenler analizi veri setinin boyutunu indirgeme metodu olarak kullanılmaktadır. Bunun sebebi veri setini yeniden ifade etmenin boyut indirgemeye izin vermesidir. Temel bileşenler analizi, bir veri setinin varyans-kovaryans yapısını, bu değişkenlerin doğrusal birleşimleri yardımıyla açıklayarak, boyut indirgenmesi ve yorumlanmasını sağlayan çok değişkenli bir istatistik yöntemdir [3].

Bir veri setinde p sayıda değişken varsa bu değişkenlerin belirlediği toplam değişkenliği ifade etmek üzere k sayıda temel bileşen bulmak, daha az sayıda değişken ile çalışarak p boyutlu bir veri seti yerine k boyutlu ($k<p$) bir veri setinde çalışmak ve böylece boyut indirgemek amaçlanır.

Eğer bir uygulamada kullanılmak üzere aralarında korelâsyon bulunan p değişkene ilişkin veri toplanmış ise toplam değişkenliği ifade etmek üzere aralarında korelâsyon bulunmayan k sayıda doğrusal bileşen ile de oluşumu açıklamak mümkündür. Temel bileşenler analizinde, karşılıklı bağımlılık gösteren, p adet değişken; doğrusal ve birbirinden bağımsız olma özellikleri taşıyan k tane yeni değişkene dönüştürülmektedir.

Veri setini korelâsyonlardan bir şekilde arındırarak kullanmak gerekir. P sayıda değişkeni, bu değişkenlerin doğrusal bileşenleri olan ve aralarında korelâsyon bulunmayan k sayıda yeni yapay değişkenlerle ifade etmek mümkündür. Bu işlem temel bileşenler analizi kullanılarak yapılabilmektedir.

Temel bileşenler analizinde amaç tek bir grup halinde bulunan çok sayıdaki değişkenlerin boyut indirgeyerek anlamlı ve daha kolay açıklanabilir daha az sayıdaki değişkenle ifade etmektir. Aynı zamanda aralarında yüksek düzeyde korelâsyon bulunan verilerden daha az sayıda ve aralarında korelâsyon bulunmayan değişkenler türetmek ve boyut indirgemesi amacıyla uygulanılan bir yöntemdir [3].

Temel bileşenler analizinin genel amacı veri indirgeme ve yorumlamadır. Sistemdeki toplam değişkenliğin tamamını oluşturmak için, p tane gerekli bileşen olmasına rağmen, bu değişkenliğin mümkün olduğu kadar büyük bir kısmı k<p olmak üzere k tane temel bileşen ile açıklanabilir. Bu durumda k tane ana bileşende hemen hemen orijinal p tane değişkende olduğu kadar bilgi mevcuttur. k tane temel bileşen başlangıçta p tane değişken ile yer değiştirir ve p tane değişkenin n adet gözleminde oluşan bir veri setine indirgenmiş olur.

Ele alınan çalışmada çok sayıda değişken ile çalışılmakta ise bu değişkenlerin hepsini bir arada anlamlı bir şekilde açıklamak zordur. Bu sebepten boyut indirgeme metodu kullanılmaktadır. Boyut indirgeme yöntemlerinden biri olan temel bileşenler analizinde orijinal değişkenlere dönüşümler uygulanarak yeni temel bileşenler elde edilir.

2.2.1. Temel bileşenler analizinin özellikleri

Temel bileşenler analizi çoklu ilişki problemini ortadan kaldırmak amacı ile kullanılan bir tekniktir.

Temel bileşenler, değişkenlerin doğrusal bileşenleri olarak değerlendirildikleri halde kendi başlarına değil, dolaylı olarak ölçülürler. Böylece veri setinde farklı bir bakış açısı sağlayıp, yorumlama ve analizde kolaylık sağlamaktadır.

Temel bileşenler analizinde önemli olan hususlardan bazıları şunlardır:

- Temel bileşenler analizi sonucu ortaya çıkan her bir yeni değişken, orijinal değişkenlerin doğrusal bir birleşimidir.
- Birinci temel bileşen, verilerdeki en yüksek varyansı açıklayacak ve baskınlığı sağlayacak şekilde türetilmektedir.

- Geriye kalan bileşenler sırasıyla en yüksek varyansı sağlarlar. Yani birinci temel bileşen en çok, diğer bileşenler ise gittikçe azalan miktarlarda toplam varyansa katkıda bulunurlar. Bu nedenle, az sayıda bileşenle toplam varyansın büyük bir kısmı açıklanabilmektedir. En fazla orijinal değişken sayısı kadar temel bileşen üretilebilmektedir.
- Yeni bileşenler birbirinden bağımsız olarak türetilmektedir.

Çoklu regresyon analizi uygulanması istenilen fakat regresyon varsayımlarından birisi olan çoklu bağımlılık koşulunun yerine gelmemesi nedeniyle regresyon analizinin uygulanamadığı durumlarda veriler temel bileşenlere göre temel bileşen skorlarına dönüştürülür ve yeni elde edilen verilere çoklu regresyon uygulanır.

2.2.2. Temel bileşenlerin korelasyon (standartlaştırılmış değişkenler) matrisinden elde edilmesi

Standartlaştırma işlemlerinde genel olarak kullanılan yöntem, her bir değişkeni o değişkenin sütun ortalamasından çıkartıp standart sapmasına bölerek varyansları birleştirip küçültür ve korelasyon matrisinde işlem yapmayı gerektirir.

Temel bileşenler analizi uygulanacak veri setindeki değişkenler farklı birimlerde ölçülmüş olabilirler. Veri setindeki değişkenleri aynı formatta ifade etmek gerektiği için standartlaştırma işlemi uygulanır. Bu standartlaştırma işlemi uygulanmış değişkenlere de temel bilişenler analizi uygulanabilir.

Bunlar,

$$z_1 = \frac{(X_1 - \mu_1)}{\sqrt{\sigma}_{11}} \qquad (2.1)$$

$$z_2 = \frac{(X_2 - \mu_2)}{\sqrt{\sigma}_{22}} \quad .$$

.

.

.

$$z_n = \frac{(X_n - \mu_n)}{\sqrt{\sigma}_{nn}}$$

şeklinde gösterilir.

Matris notasyonuyla ise,

$$Z = \left(V^{\frac{1}{2}}\right)^{-1} (X - \mu) \qquad (2.2)$$

şeklinde gösterilmektedir.

Burada, köşegen standart sapma matrisi $V^{\frac{1}{2}}$,

$$V^{\frac{1}{2}} = \begin{bmatrix} \sqrt{\sigma}_{11} & \cdots & 0 \\ 0 & & 0 \\ \vdots & \ddots & \vdots \\ \vdots & & \vdots \\ 0 & \cdots & \sqrt{\sigma}_{nn} \end{bmatrix} \qquad (2.3)$$

biçiminde tanımlanmıştır.

2.3. Temel Bileşenler Analizinde Özdeğer ve Özvektörlerin Bulunması

A bir n × n boyutlu kare matris olsun. Eğer λ bir skaler ve x vektörü de sıfır olmayan, $x \neq 0$, bir sütun vektörü olmak üzere,

$$Ax = \lambda x \qquad (2.4)$$

eşitliği sağlanıyorsa x vektörü, A matrisinin özvektörü, λ skaleri de A matrisinin özdeğeridir. Aynı zamanda x, λ özdeğerine karşılık gelen özvektördür.

Bir skaler olan λ, n × n boyutlu A matrisi için Ax = λx denkleminde x'in sonsuz çözümü olduğu durumda bir özdeğer tanımlar [5].

Temel Özellikler

- Özdeğer λ sıfır değerini alabilirken, özvektör x asla sıfır vektörü olamaz.
- Özdeğer sıfır olduğunda Ax = 0x, A matrisinin tersi alınamaz.
- Boyutu n×n olan bir A matrisinin tersinin alınabilir olması için tüm özdeğerlerinin sıfırdan farklı olması gerekir.

2.3.1. Analizde temel bileşenlerin seçimi ve sayısının belirlenmesi

Temel bileşenler hesaplanırken özdeğerlerin bulunmasından sonra önemli özdeğer sayısına karar vermek çok önemlidir. Bu amaçla birçok yöntem geliştirilmiştir. Bu yöntemlerden biri, standartlaştırılmış veri matrislerinin kullanılması halinde birden büyük değerli özdeğerlerin sayısı m sayısını vermektedir veya yaklaşık aynı mantığa dayanan,

$$\sum_{j=1}^{m} \frac{\lambda_j}{p} \geq \frac{2}{3} \tag{2.5}$$

koşulunun sağlandığı en küçük m değeri önemli temel bileşenlerin sayısı olarak alınmaktadır.

3. BÜYÜK VERİLERDE REGRESYON ANALİZİ

3.1. Doğrusal Regresyon Analizi

Regresyon analizi, aralarında sebep-sonuç ilişkisi bulunan iki veya daha fazla değişken arasındaki ilişkiyi belirlemek ve bu ilişkiyi kullanarak o konu ile ilgili tahminler (estimation) ya da kestirimler (prediction) yapabilmek amacıyla yapılır.

Sebep ve sonuç olarak örnek vermek gerekirse sırasıyla gelir-harcama, yaş-boy, gübre-verim verilebilir.

Bu analiz tekniğinde iki (basit regresyon – tek değişkenli) veya daha fazla değişken (çoklu regresyon – çok değişkenli) arasındaki ilişki açıklamak için matematiksel bir model kullanılır ve bu model regresyon modeli olarak adlandırılır.

3.1.1. Tek değişkenli regresyon analizi

Tek değişkenli regresyon analizi bir bağımlı değişken ve bir bağımsız değişken arasındaki ilişkiyi inceleyen analiz tekniğidir. Bu analizle bağımlı ve bağımsız değişkenler arsındaki doğrusal (lineer) ilişkiyi temsil eden bir doğru denklemi formüle edilmektedir. Korelasyon analizinde olduğu gibi regresyon analizinde üzerinde durulan ilişki,değişkenler arasındaki doğrusal ilişkidir. Bu doğrunun hesaplanması ise en küçük kareler metodu yardımıyla yapılmaktadır[6].

3.1.2. Çok değişkenli regresyon analizi

Çok değişkenli regresyon analizi, bir bağımlı değişken ve birden fazla bağımsız değişkenin yer aldığı regresyon modellerine denir. Çok değişkenli regresyon analizinde bağımsız değişkenler aynı anda bağımlı değişkendeki değişimi açıklamaya çalışmaktadır. Çok değişkenli regresyon analizinin yorumu bazı farklılıklar olması ile birlikte tek değişkenli regresyon analizine benzemektedir.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \quad (3.1)$$

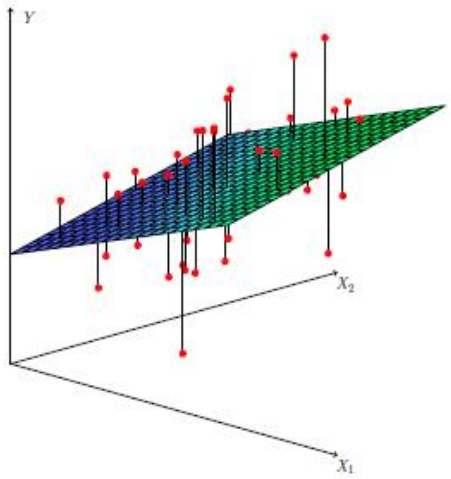
y: Bağımlı değişken

X değerleri: Bağımsız değişkenler

β değerleri: Her bir X değerine karşılık gelen ağırlıklar

3.1.2.1. Regresyon analizinin en küçük kareler yöntemi ile elde edilmesi

En küçük kareler yöntemi, oluşturulan regresyon modelinde X değerlerine karşılık gelen y değerlerinin bulunduğu grafikte, (Xi , yi) veri noktalarının oluşturulan modele uzaklıklarının karelerinin minimize edilmesini amaçlamaktadır.



Şekil 3.1. Veri noktalarının oluşturulan modele en küçük kareler yöntemi için uzaklıklarının gösterilmesi

$$f(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \quad (3.2)$$

Yukarıdaki formül üzerinden en küçük kareler yöntemi kullanarak elde edilen hata kaybını minimize etmek istenir ise elde edilecek maliyet fonksiyonu R:

$$R = \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2 \quad (3.3)$$

Burada her bir gerçek y değerinden tahmin edilen y değeri ($f(x_i)$) çıkarılıp karelerinin toplamı alınarak toplam hata minimize edilmiş olur. Burada dikkat edilmesi gereken nokta her bir gerçek y değerinden çıkarılan, tahmin edilen y($f(x_i)$) değerinin karelerinin neden alındığıdır. Karelerinin toplamının alınmasındaki amaç elde edilebilecek negatif değerlerin toplam hatanın yanlış bulunmasını sağlayacak olmasıdır. Bu da modelimizin doğruluktan sapacağı gerçeğini göstermektedir. En küçük kareler yöntemi ile her bir farkın karelerinin toplamı alınmış olduğu için doğru sonuç elde edilmiş olur.

Minimize etmek istenilen denklem ($y_i = \beta_0 + \beta_1 X_i$ *olduğunu varsayalım*) yukarıda elde edilen denklem(R) ile birlikte kullanıldığında:

$$R(\beta) = \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 X_i\right)^2 \quad (3.4)$$

denklemi elde edilmiş olur.

Burada R değerini minimum yapan β_0 *ve* β_1 değerlerini bulalım. Bunun için R nin β_0 *ve* β_1 e göre kısmi türevlerini 0 yapan değer bulunmalıdır[7].

β_0 için:

$$\frac{\partial R}{\partial \beta_0} = 2 \times \frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 X_i) \times \frac{\partial}{\partial \beta_0}(y_i - \beta_0 - \beta_1 X_i) \quad (3.5)$$

$$\frac{\partial R}{\partial \beta_0} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 X_i) \times (-1) = 0 \quad (3.6)$$

$$\beta_0 = \frac{1}{n}\sum_{i=1}^{n}(y_i) - \beta_1 \frac{1}{n}\sum_{i=1}^{n}(X_i) \quad (3.7)$$

β_1 için:

$$\frac{\partial R}{\partial \beta_1} = 2 \times \frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 X_i) \times \frac{\partial}{\partial \beta_1}(y_i - \beta_0 - \beta_1 X_i) \quad (3.8)$$

$$\frac{\partial R}{\partial \beta_1} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 X_i\right) \times (-X_i) = 0 \qquad (3.9)$$

$$\beta_1 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i{}^2\right) = \sum_{i=1}^{n}\left(X_i\, y_i\right) - \sum_{i=1}^{n}(\beta_0 X_i) \qquad (3.10)$$

β_0 değerini β_1 değerinin içine yazacak olur isek:

$$\beta_1 = \frac{\sum_{i=1}^{n} X_i\, y_i - \frac{1}{n}\sum_{i=1}^{n} y_i \sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i{}^2 - \frac{1}{n}\sum_{i=1}^{n} X_i \sum_{i=1}^{n} X_i} \qquad (3.11)$$

şeklinde elde edilmiş olur.

Birden fazla değişken için bu işlem gerçekleştirilmek istenir ise:

$$f(x) = \beta_0 + \sum_{j=1}^{n}\beta_j X_j \qquad (3.12)$$

$$R = \frac{1}{2n}\left(y_i - \beta_0 - \sum_{j=1}^{n}\beta_j X_{ij}\right)^2 \qquad (3.13)$$

Burada bulunan β değerlerinin modelimize etkisinden bahsedecek olursak her bir β değeri çarpanı olduğu bağımsız değişken değeri açısından bakılacak olur ise 1 birimlik artış sağlandığı kadar modele etki etmiş olur. Örnek olarak $y = \beta_0 + \beta_1 X_1$ olsun. Buradaki β_0 değerine 5, β_1 değerine 10 ve X_1 değerine de 15 diyelim. X_1 değeri her bir birimlik artış gösterdiğinde y değeri de her bir birimlik artışda β_1 değeri kadar artış sağlamış olacaktır.

3.2. Doğrusal Olmayan Regresyon Analizi

Doğrusal olmayan regresyon analizi, gözlemsel verilerin model parametrelerinin doğrusal olmayan bir kombinasyonu olan ve bir veya daha fazla bağımsız değişkene bağlı olan bir fonksiyonla modellendiği bir regresyon analizi şeklidir .

Doğrusal olmayan regresyon modeli,

$$y_n = f\left(x_{n,}\theta_*\right) + \epsilon_n \quad n = 1, 2, 3 \ldots, n \qquad (3.14)$$

olarak tanımlanmaktadır.

Parametrelerden bahsedecek olur isek,

- ϵ_n = Sezgisel Hata Terimi
- x_n = Bağımsız Değişkenler
- θ = p tane bilinmeyen ve doğrusal olmayan parametre
- $f(x_n, \theta_*)$: regresyon modeli

anlamına gelmektedir[8].

Bir modelde bağımsız değişkenler ile bağımlı değişkenler arasında doğrusal bir model olmadığında doğrusal olmayan yöntemler dikkate alınabilmektedir.
Burada kullanılabilen yöntemlerden bir tanesi model üzerinde dönüşüm yaparak doğrusal bir modele benzetmektir. Burada dönüştürülen veri (logaritmik dönüşüm vs.) doğrusal bir analizle tahmin edilir ve sonuçlar incelenir[9].
Kullanılan yöntemlerden bir diğeri ise doğrusal olmayan tahmin yöntemleri kullanarak parametreleri tahmin etmektir.
Verilen bir doğrusal olmayan modelin parametrelerinin tahmin edilebilmesi için kulanılan yöntemlerden bir tanesi en küçük kareler yöntemidir.
En küçük kareler yöntemi verilen herhangi bir θ değeri için r_n artık değerler olma üzere $r_n = y_n - f(x_n, \theta)$ doğrusal olmayan model için en küçük kareler toplamı,

$$S(\theta) = \sum_{i=1}^{n} [y_i - f(x_i, \theta)]^2 \qquad (3.15)$$

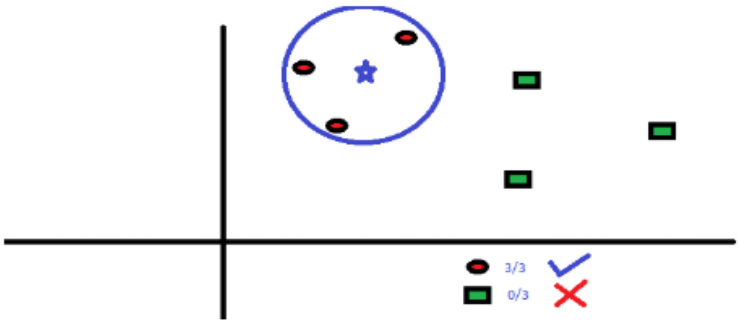
şeklinde ifade edilmektedir.
Burada en küçük kareler tahmini $\hat{\theta}$ ile gösterilir ve $S(\theta)$ en küçük kareler toplamı fonksiyonu en küçük hata oranını içerecek duruma getirmiş olur.

4. BÜYÜK VERİLERDE SINIFLANDIRMA

4.1. K En Yakın Komşu Algoritması Kullanılarak Sınıflandırma

Sınıflandırma, bir veri kümesinde bulunan değerlerin hangi sınıfa ait olduğunun bulunmasında kullanılmaktadır. Sınıflandırma algoritmalarında sistem veri setini eğitim verisi üzerinde işlem yaparak öğrenir. Daha sonra sistemin hiç bilmediği bir test verisi üzerinde sınıfı bilinmeyen değerlerin sınıfının bulunmasında kullanılır. Örnek olarak bu projede çeşitli özellikler aracılığı ile belirtilen yerlerde orman yangını olup olmadığı test edilmiştir.

En yakın komşular (KNN) algoritması, hem sınıflandırma hem de regresyon sorunlarını çözmek için kullanılabilecek basit, uygulaması kolay bir makine öğrenmesi algoritmasıdır. KNN algoritması içerisinde bulunan k değeri, yeni gelen bir veri noktasının kendisine en yakın kaç komşuya bakılması gerektiğini belirtir. Örnek olarak k=3 dersek, bu k değeri yeni gelen veri noktasının kendisine en yakın 3 komşusuna bakması gerektiğini göstermektedir[10].



Şekil 4.1. En Yakın Komşu Gösterimi Grafiği

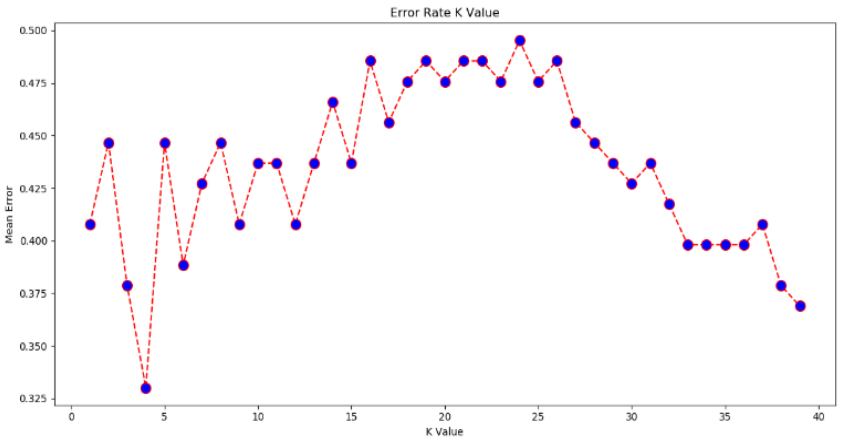
Yukarıdaki grafikte bulunan kırmızı noktalar bir sınıfı, yeşil noktalar da bir sınıfı göstermiş olsun. Yıldız ile gösterilen ve yeni gelmiş bir veri noktasının en yakın 3 komşusuna bakacak olursak 3 komşu da kırmızı noktalardan oluştuğu için yeni gelen veri noktamız da kırmızı sınıfta bulunduğu kabul

edilir. Burada bulunan noktalar sınıflandırma yapıldığında seçilen komşu sayısı göz önüne alınarak en çok hangi sınıftan veri bulunuyorsa yeni veri o sınıfta yer almış olur. Regresyon problemi için bakacak olur isek seçilen komşu sayısı göz önüne alındığında komşuların değerlerinin toplanıp komşu sayısına bölündüğünde yeni gelen verinin değeri bulunmuş olur.

KNN algoritması noktalar arası mesafeyi belirlemek için çeşitli matematiksel metrikler kullanır. Bu metriklere örnek olarak Öklid ve Minkowski mesafeleri verilebilir. Bu projede Öklid mesafesi kullanıldığı için nasıl hesaplandığına değinelim. N boyutlu öklid uzayında $P = (p_1, p_2, \ldots., p_n)$ ve $Q = (q_1, q_2, \ldots., q_n)$ noktaları arasındaki öklid uzaklığı şu şekilde tanımlanır:

$$\sqrt{\sum_{i=1}^{n}(p_1 - q_1)^2} \tag{4.1}$$

KNN algoritmasının ihtiyaç duyduğu bir başka özellik ise kullanılan veriye uygun en iyi sonucu veren k değerinin belirlenmesidir. Çeşitli k değerleri veri üzerinde denenerek uygun k değerine yaklaşılabilir. Ayrıca belirli bir k değeri ve bu k değerlerine karşılık gelen doğruluk değerleri veya ortalama hata değerleri grafiksel gösterilerek en uygun k değeri daha net bir şekilde gösterilebilir [11].



Şekil 4.2. En Yakın Komşu Değerlerine Göre Ortalama Hata Grafiği

Yukarıdaki grafikte görüldüğü üzere 1 ile 40 arasında bulunan k komşu değerlerine karşılık gelen ortalama hata değerlerinin grafiği gösterilmektedir. Burada uygulanan veri üzerinde en az ortalama hatayı veren komşu değerinin 4 olduğu görülmektedir ve test bu değer üzerinden yapılmalıdır.

KNN algoritmasının avantajlarından bahsedecek olursak basit ve uygulaması kolay bir algositmadır. Bunun yanı sıra model oluşturmaya, birkaç parametre ayarlamaya veya ek varsayımlarda bulunmaya gerek yoktur. Ayrıca KNN algoritması çok yönlüdür. Sınıflandırma, regresyon ve arama için kullanılabilir.

KNN algoritması sınıflandırma problemleri için uygun sonuçlar verebilmekle beraber özellik sayısı arttığında çok iyi sonuçlar verememektedir. Bunun nedeni KNN algoritmasının verilen k değerine göre ve verilen bağımsız değişken sayısı arttıkça komşuları ararken tahminlerin hızla yapılması gereken durumlarda pratik olmayan bir seçim olmasını sağlar.

4.2. Lojistik Regresyon

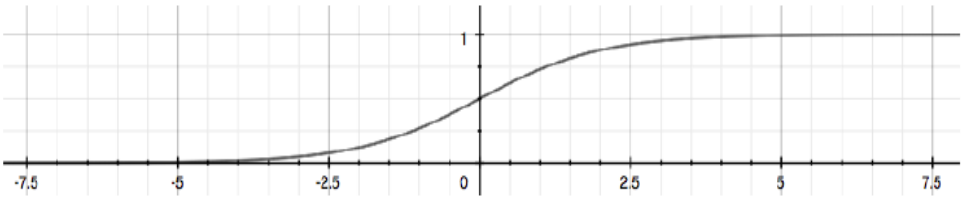
Lojistik regresyon, bağımlı değişken sayısı ikili(binary) olduğunda yapılması gereken uygun regresyon analizidir. Diğer bütün regresyon analizleri gibi tahminsel bir analiz şeklidir. Verileri tanımlamak ve bir bağımlı ikili değişken ile bir veya daha fazla bağımsız değişkenler arasındaki ilişkiyi açıklamak için lojistik regresyon kullanılır.

Lojistik regresyon gerçekte bir regresyon problemi değil sınıflandırma problemlerine bir yaklaşımdır ve gerçek regresyon problemleriyle karıştırılmamalıdır. Çıktı değerimiz olan değer (y) sürekli devam eden bir değer aralığında olması yerine sadece 0 veya 1 arasında olmalıdır.

$$y \in \{0,1\} \tag{4.2}$$

Burada bulunan 0'ın genellikle negatif sınıfı içerdiği ve 1'in ise pozitif sınıf olarak atandığı görülmektedir. Bir metot doğrusal regresyon kullanarak 0.5 ten daha küçük ve sıfıra kadar olan değerlere 0, 0.5 ten büyük ve 1 e kadar olan değerlere 1 vermektedir.

Burada bulunan y bağımlı değişkenlerinin 0 ile 1 arasında olmasını sağlayan lojistik fonksiyona sigmoid fonksiyonu adı verilir.



Şekil 4.3. Sigmoid Aktivasyon Fonksiyonu

Şekilde bulunan sigmoid fonksiyonu için bağımlı değişkenden gelen gerçek değerleri 0 ile 1 arasına almamızı sağlayan denklem

$$g(z) = \frac{1}{1+e^{-z}} \tag{4.3}$$

şeklinde gösterilmektedir. Burada bulunan g(z) x e bağlı olarak y nin gerçekleşip gerçekleşmeme olasılığı, z değeri ise doğrusal regresyondan da bildiğimiz üzere doğrusal regresyon denklemiyle aynıdır.

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \tag{4.4}$$

Burada doğrusal regresyon denkleminden gelen değer sigmoid fonksiyon denklemi olan g(z) denklemine uygulandığında çıkan sonuç 0 ile 1 arasında olmaktadır. Bu g(z) denkleminden gelen sonuç bize çıktımızın 1 olma olasılığını vermektedir. Örnek olarak denklemden gelen değer 0.7 ise çıktımızın %70 olasılıkla 1 olması gerektiğini göstermektedir. Aynı şekilde gelen değerin 0 olma olasılığı da 1 olma olasılığının tamamlayıcısıdır. Yani 0.7 değerinin 1 olma olasılığı %70 ise 0 olma olasılığı da %30 dur.

Sınıflandırmamızdan gelen değer 0 ile 1 arasına alındığı için bu değerlerin 0 mı yoksa 1 mi olduğuna karar vermek yani düzgün bir şekilde sınıflandırma yapabilmek için lojistik fonksiyondan gelen değer için bir eşik değeri (threshold) belirlenmesi gerekmektedir. Bu eşik değeri genel olarak 0.5 ten büyük ise 1, 0.5 ten küçük ise 0 olarak belirlenmektedir.

$$g(z) \geq 0.5 \rightarrow y = 1 \tag{4.5}$$

$$g(z) < 0.5 \rightarrow y = 0 \tag{4.6}$$

Burada bulunan g(z) denklemi doğrusal olmadığı için z denklemi içerisinde bulunan β parametrelerini tahmin etmek zordur. Bu g(z) denklemi doğrusallaştırıldığında

$$L_i = \ln\left(\frac{g(z)}{1-g(z)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \tag{4.7}$$

elde edilen fonksiyona lojit fonksiyonu denir ve böylece katsayılar kolayca elde edilebilir[12].

4.2.1. Lojistik regresyon katsayısının yorumlanması

Lojistik regresyonda parametrelerin yorumlanması doğrusal regresyonda bulunan parametrelerin yorumlanması kadar kolay değildir. Doğrusal regresyonda x de gerçekleşecek bir birimlik artışta y değeri x değeri kadar artış veya azalış göstermekteydi. Lojistik regresyonda ise x deki bir birimlik artış için $\frac{g(z)}{1-g(z)}$ odds tahmin formülünden gelen değer ile exp(β) değerinin çarpımından elde edilen lojistik fonksiyonundan faydalanılır[13].

5. UYGULAMALAR

Bu uygulamada büyük verilerde doğrusal ve doğrusal olmayan regresyon analizi yardımıyla orman yangını tahmini yapılması amaçlanmıştır. Projede regresyon analizi yapılmadan önce regresyon analizine veri hazırlamak ve boyut indirgemek için temel bileşen analizi yapılmıştır. Doğrusal ve doğrusal olmayan regresyon analizi yapabilmek için bir orman yangınları verisi bulunmuştur. Bu veri seti 517 satır ve 13 sütundan oluşmaktadır.

5.1. Veri Seti

Sütun özellikleri sırasıyla,

- X – Montesinho park haritası içerisindeki x ekseni uzaysal koordinatı: (1 ile 9 arası)
- Y – Montesinho park haritası içerisindeki y ekseni uzaysal koordinatı (2 ile 9 arası)
- Ay – Yangınların hangi aylarda çıktığı bilgisini tutmaktadır.
- Gün – Yangınların hangi günlerde çıktıklarının bilgisini tutmaktadır.
- FFMC – İnce yakıt nemi kodu, çöp ve diğer sertleşmiş ince yakıtların nem içeriğinin sayısal bir derecesidir. Bu kod, bağıl tutuşma kolaylığının ve ince yakıtın yanıcılığının bir göstergesidir.(18.7 ile 96.20 arası)
- DMC – Duff nem kodu, orta derecede derinlikte gevşek biçimde sıkıştırılmış organik tabakaların ortalama nem içeriğinin sayısal bir derecesidir. Bu kod, orta kat katlarında ve orta büyüklükte odunsu malzemede yakıt tüketiminin bir göstergesidir.(1.1 ile 291.3 arası)
- DC – Kuraklık kodu, derin, kompakt organik tabakaların ortalama nem içeriğinin sayısal bir derecesidir. Bu kod, mevsimsel kuraklığın orman

yakıtları üzerindeki etkisini ve büyük kütüklerdeki yanma için etkili bir göstergedir.(7.9 ile 860.6 arası)[14].

- ISI - İlk Yayılma Endeksi, beklenen yangın yayılma oranının sayısal bir değerlendirmesidir. Değişken miktarlarda yakıtın etkisi olmadan rüzgarın ve FFMC'nin yayılma hızı üzerindeki etkilerini birleştirir. (0 ile 56,10 arası)
- Sıcaklık – Sıcaklık santigrat derece cinsinden gösterilmektedir. (2.2 ile 33.30 arası)
- RH – Yüzde olarak bağıl nem miktarını göstermektedir. (15 ile 100 arası)
- Rüzgar – km / saate göre rüzgar hızını göstermektedir. (0.40 ile 9.40 arası)
- Yağmur – mm / m2 oranına göre dışarıda bulunan yağmur miktarını göstermektedir. (0.0 ile 6.4 arası)
- Alan – Ormanın yanık alanı ha cinsinden. (0 ile 1090,84 arası)

şeklinde belirtilmiştir.

Veri setimizin içerisinde nümerik olmayan değerler bulunmaktadır. Temel bileşenler analizi sadece nümerik değerler ile yapılabilmektedir. Bu nedenden dolayı veri seti içerisinde bulunan ay ve gün değerleri sayısal karşılıkları olan değerlere çevrilmiştir. Bu işlem gerçekleştirildikten sonra veri seti içerisinde bulunan bütün değerler nümerik değere çevirilmiş şekilde kullanıma hazırlanmıştır. Aynı zamanda veri seti içerisinde hiçbir eksik değer (missing value) bulunmamaktadır.

Veri setine temel bileşenler analizi uygulanarak veri seti içerisinde bulunan en baskın, birbirinden bağımsız, korelasyonsuz ve veri kaybı en az olan 3 sütun değeri elde edilmiştir. Bu sütun değerlerinini, diğer bütün sütun değerlerine göre oranlandığında alınan sonuç yüzde 99.51 olarak gözlemlenmiştir. Bu da çok çok az, göz ardı edilebilecek bir kayıpla temel bileşenler analizinin tamamlanmış olduğunu göstermektedir. Bunun yanı sıra temel bileşenler analizinin veri setimiz için en doğru analiz tekniği olduğu da söylenebilmektedir.

Temel bileşenler analizi sonucunda elde edilmiş 3 en baskın sütun değerleri ise,

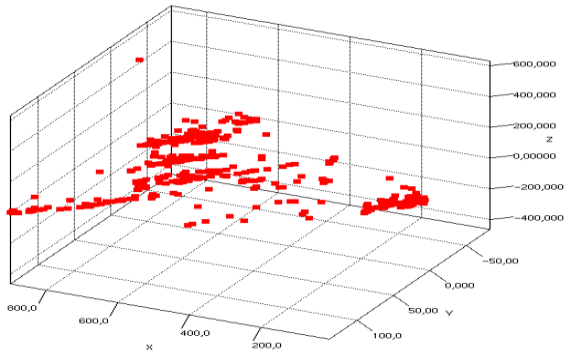
- DC – Kuraklık kodu
- Area (Alan) – Yanmış alan miktarı
- DMC – Sıkıştırılmış organik tabakaların ortalama nem içeriği

şeklinde gösterilmektedir.

Temel bileşen analizi ile elde edilen bu 3 değişken ile alınacak verim, analiz ile indirgenmiş diğer 10 değişken ile alınacak verim kadar olacaktır. Elde edilen kazanç, daha az veri ile çalışarak aynı verimin alınacak olmasıdır.

Değişken değerlerinin ayrı ayrı etkisini görecek olursak Kuraklık kodu (DC) yüzde 90.69 ile en çok etkiyi gösteren değişken olurken , Alan (Area) yanmış alan miktarı yüzde 5.78 ile ve DMC nem içeriği yüzde 3.02 ile etki etmektedir. Burada kuraklık kodunun bütün veri seti arasında çok büyük bir etkisinin olduğu görülmektedir.

Son olarak yapılan bir gözlemde ise veri setimizde bulunan noktaların grafiksel dağılımının gözlenmesidir. Veri setinde bulunan herbir nokta için scatter plot grafiği çizdirilmiştir.



Şekil 5.1. Veri noktaları grafiği (Scatter plot)

Temel bileşenler analizi sonucunda çıkartılan 3 değişkenin veri noktalarının dağılım grafiği şekilde görüldüğü gibidir.

Veri setini düzeltme işlemi tamamlandıktan sonra veri seti üzerine z-score normalizasyon işlemi uygulanmıştır. Z-Score normalizasyon işlemi her bir veri noktasını bulunduğu sütunun ortalamasından çıkartıp yine bulunduğu sütunun standart sapmasına bölerek gerçekleştirilmektedir. Z-Score normalizasyon işleminin yapılmasının nedeni baskın değerleri ortadan kaldırmaktır. Veri seti

içerisinde bulunan herbir sütunu ve veri noktasını birbiri cinsinden kıyaslayabilmek gerekmektedir. Aksi taktirde baskın değerler sonucu çok farklı yerlere çekebilmektedir. Bu yüzden verileri belli bir aralığa almak ve birbiri cinsinden yazmak için normalizasyon işleminin formülü içerisinde bulunan standart sapmaya bölme işlemi gerçekleştirilir.

Z-Score normalizasyon işlemi,

$$z_1 = \frac{(X_1 - \mu_1)}{\sqrt{\sigma}_{11}} \tag{5.1}$$

$$z_2 = \frac{(X_2 - \mu_2)}{\sqrt{\sigma}_{22}}$$

.

.

.

$$z_n = \frac{(X_n - \mu_n)}{\sqrt{\sigma}_{nn}}$$

şeklinde formulüze edilmektedir.

Veri setimizi normalize ettikten sonra kovaryans matrisi elde etmemiz gerekmektedir. Kovaryans kavramı değişkenler arası doğrusal ilişkinin değişkenliğini ölçen bir kavramdır. Sonucun pozitif olması artan bir doğrusal ilişkiyi, negatif olması azalan bir ilişkiyi ve sıfır olması bir ilişkinin bulunmadığını göstermektedir.

Kovaryans matrisi ise bu değişkenlerin karşılıklı kovaryans değerlerinin bulunduğu bir matristir. Kovaryans matrisi, uygulamamızda normalize edilmiş veri seti matrisinin yine bu matrisin transpozu ile çarpılarak elde edilmiştir.

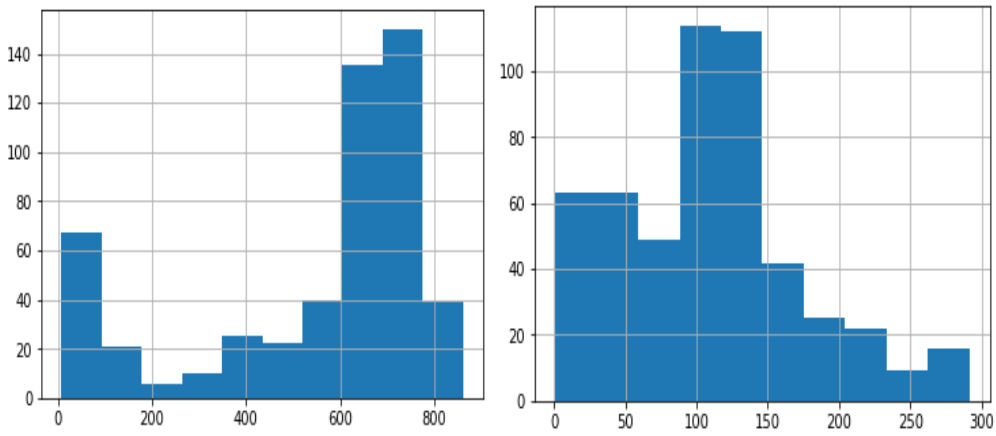
Normalize edilmiş matris ile transpozu çarpılarak elde edilen kovaryans matris bir kare matristir. Veri seti üzerine temel bileşen analizi uygulanabilmesi için ve özdeğer ve özvektörlerinin çıkartılabilmesi için matrisin kare matris olması gerekmektedir. Bu işlemler gerçekleştirildikten sonra elimizde bulunan kovaryans matris 13 x 13 lük bir kare matristir. Elde ettiğimiz kovaryans matris kullanılarak verimizin özdeğerleri ve özvektörleri bulunmuştur.

Bilindiği üzere temel bileşenler analizi boyut indirgemek ve en az veri kaybına sebep olmak için kullanılmaktadır. Özdeğer ve özvektörler çıkarıldıktan sonra en baskın 3 sütun değeri elde edilmiştir ve boyut indirgeme gerçekleştirilmiştir. Elde ettiğimiz yeni veri seti 13 satır ve 3 sütuna indirgenmiştir.

Daha sonra bu elde edilen matris orijinal veri setimizle çarpılarak 517 satır 3 sütundan oluşan birbirinden bağımsız, korelasyonsuz ve en baskın olan satır ve sütun değerleri elde edilmiştir. Son olarak projenin şuan ki kısmına kadar yapılmış olan kısmı olan grafik çizdirme işlemi gerçekleştirilmiştir.

Elimizde bulunan sütunların herbirinin bir boyut olduğunu düşünürsek 3 boyutlu bir grafik çizdirme işlemi gerçekleştirilmiştir. Herbir satırda bulunan ve 3 sütuna ait veri noktaları (x, y, z) şeklinde bir nokta haline getirilerek 3 boyutlu düzlem üzerinde scatter plot aracılığıyla gösterilmektedir.

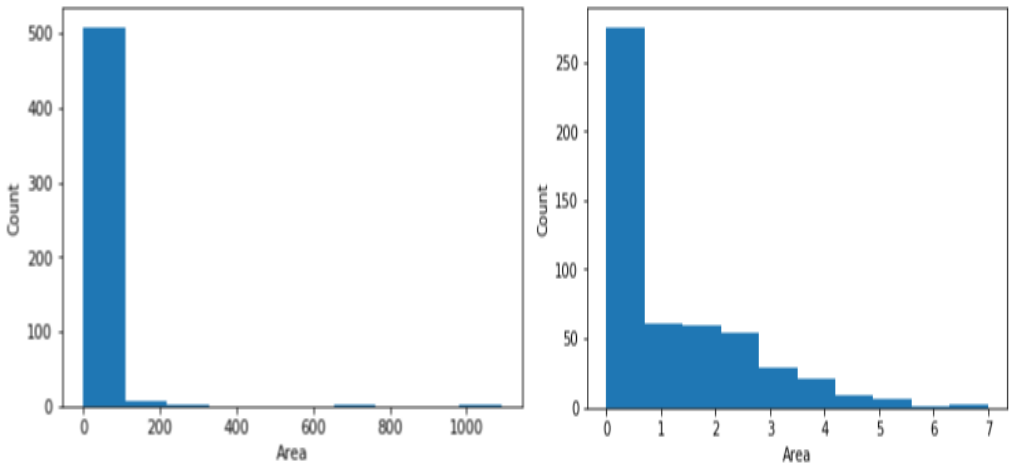
Orman yangınları veri seti çok zor bir regresyon problemidir ve doğrusal bir regresyon modeli uydurmak oldukça zor bir iştir. Veri seti içerisinde bulunan özelliklerin çıktı özelliğimiz olan AREA özelliği karşısındaki veri dağılımlarına bakacak olur isek,



Şekil 5.2. DC ve DMC özelliklerinin veri seti içerisindeki dağılımını gösteren histogram grafiği

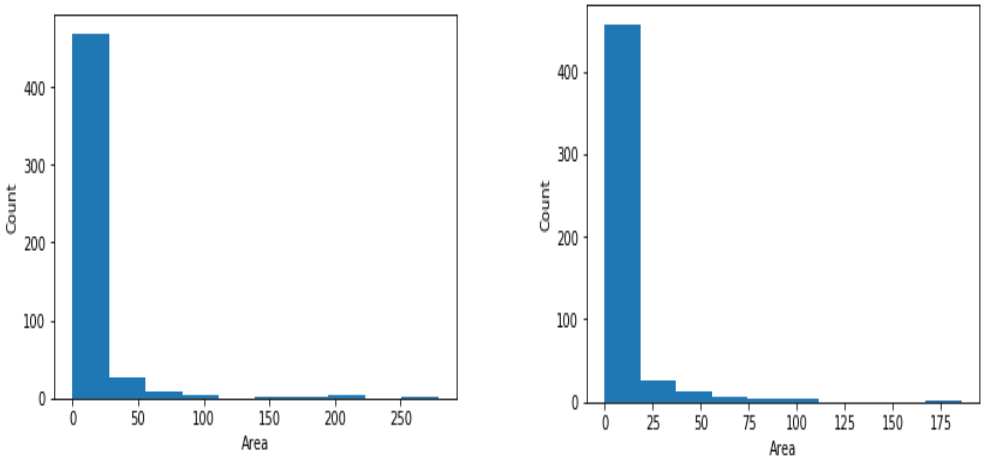
Yukarıdaki histogram grafiklerinde temel bileşenler analizinden elde edilmiş en yüksek varyansa sahip 2 özellik olan DC ve DMC özelliklerinin veri içerisindeki dağılımları görülmektedir. Buradan da görüldüğü gibi bu özellikler kendi içerilerinde normal bir şekilde dağılmamıştır. Bu nedenle regresyon eğrisinin uydurulması çok zordur.

5.2. Regresyon Eğrisinin Uydurulması



Şekil 5.3. Yanmış alan(soldaki) ve logaritmik dönüşüm yapılmış yanmış alan(sağdaki) değerlerinin bulunduğu histogram grafiği

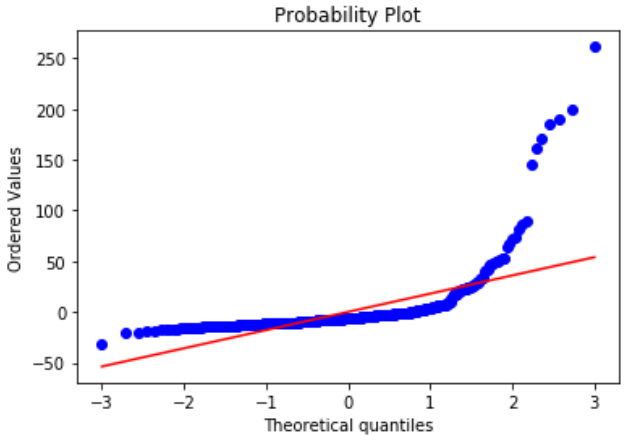
Veri setimizde bulunan verilerin dağılımından bahsedecek olursak sol tarafta bulunan grafikte veri setimizin gerçek değerlerinin dağılımı görülmektedir. Burada göze çarpan durum verilerin büyük bir çoğunluğunun 0 ve 0'ın etrafında bulunmasıdır. Bu durumdan veri setini kurtarmak için veri setine logaritmik dönüşüm (ln(area+1)) uygulanmıştır. Burada veri setinin logaritmik dönüşüm uygulanmamış kısmı kullanılarak bir regresyon modeli uydurulup uydurulamayacağının kontrolünün yapılması gerekmektedir. İşleme başlamadan önce sol tarafta bulunan grafikte göze çarpan yüksek değer denilebilecek değerlerin modelimizi bozacağı açık bir şekilde görüldüğü için yüksek görülen değerler veri seti içerisinden çıkarılmıştır.



Şekil 5.4. Veri seti içerisinden yüksek değerler çıkarıldığında veri seti içerisinde oluşan dağılımlar

Yukarıda sol tarafta bulunan grafik veri seti içerisinden yüksek değer olarak görülen veriler çıkarıldığında oluşan veri dağılımını göstermektedir. Burada gözlenen durum veri seti içerisindeki verilerin dağılımının yüksek değerler çıkarıldıkça yeni gelen veri seti içerisindeki değerlere göre yeni yüksek değerlerin oluştuğunun görüldüğüdür. Yine sağ tarafta görülen grafikte sol tarafta bulunan grafikteki yüksek değerlerin çıkarılmış halidir. Burada da görüldüğü üzere yine yeni yüksek değerler yeni veri seti için ortaya çıkmaktadır.

Yukarıda sol tarafta bulunan grafik üzerine çoklu doğrusal regresyon uygulandığında oluşan modelin grafiği aşağıda verilmektedir.

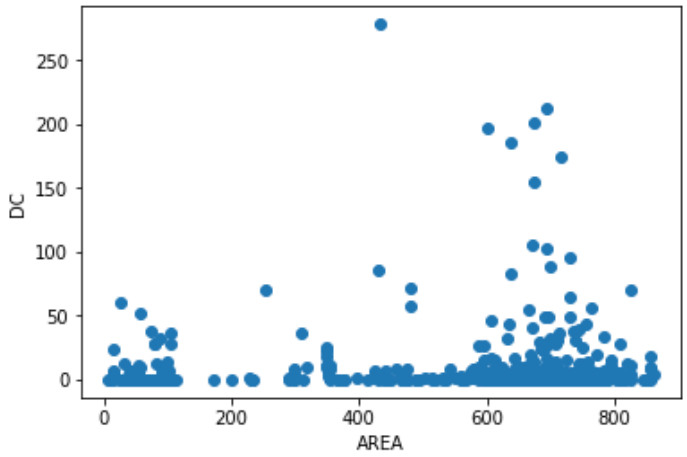


Şekil 5.5. Çoklu doğrusal regresyon uygulanan veri setinin regresyon model grafiği

Yukarıda bulunan grafikte oluşan eğriye bakıldığında modelimiz ve veri noktalarının dağılımı da göz önünde bulundurulduğunda bu veri seti üzerinde doğrusal bir eğri oluşturulamayacağı, oluşturulabilse bile sonuçların iyi olmayacağı görülmektedir. Bir veri setinde, oluşturulan eğri üzerinde gösterilen noktalar oluşan eğriye ne kadar yakın ise modelimizin doğruluğu o kadar yüksek olmaktadır. Ancak bizim oluşturduğumuz grafikte bulunan noktalar oluşan eğri ile düzgün bir uyum gösterememektedir.

Veri seti içerisinde görüldüğü üzere çok fazla 0 değeri göze çarpmaktadır. Bu sıfır değerlerinin bulunduğu satırlar model oluşturmaya olumsuz yönde etki ettiği düşünülerek veri setinden çıkartılıp yeni regresyon modeli üzerinde işlem

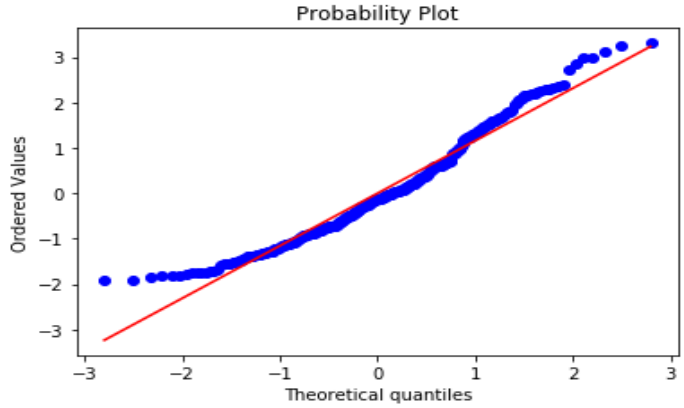
yapılıp sonuçlara bakılmıştır fakat burada da herhangi bir gözle görülen değişim bulunmamaktadır.



Şekil 5.6. En yüksek varyansa sahip özellik olan DC ile alan(AREA) çıktı özelliğinin dağılım grafiği

Yukarıdaki dağılım grafiğinde veri seti içerisinde en yüksek dağılıma sahip DC özelliği ile çıktı özelliğimiz olan AREA özelliği arasında bulunan dağılım gözükmektedir. Burada gösterilen dağılıma bakılacak olursa herhangi bir şekilde doğrusal eğri oluşturulamayacağı gözükmektedir.

Şekil 5.3. de sağ tarafta bulunan logaritmik dönüşüm uygulanmış veri üzerine çoklu regresyon modeli uygulandığında regresyon modelimiz,



Şekil 5.7. Çoklu doğrusal regresyon uygulanan logaritmik dönüşüm uygulanmış veri setinin regresyon model grafiği

şekilde görüldüğü gibi oluşmaktadır. Burada da herhangi bir doğrusal model uyumundan bahsedemeyiz. Ancak logaritmik dönüşüm yapılmamış veri setinden

daha iyi bir model ortaya çıktığı söylenebilir. Fakat burada model doğruluk sonuçları karşılaştırıldığında herhangi bir şekilde göze çarpan bir sonuç gözükmemektedir.

Bir diğer kontrol edilmesi gereken durum ise veri seti üzerine temel bileşen analizi yapılmış veri üzerinde çoklu doğrusal regresyon uygulanmasıdır. Temel bileşenler analizinde veri seti içerisinde en yüksek varyansa sahip özellik seçiminde bulunmuştuk. Bu veri seti üzerinde doğrusal regresyon uygulandığında sonuçlarda herhangi bir değişim görülmemektedir. Sonuç değişikliğinin görülmemesinin en önemli sebeplerinden bir tanesi, temel bileşenler analizi veriler arasındaki dik uzaklıklar arasındaki fark ile elde edilirken regresyon analizi, yatay eğri boyunca uzaklıkların farkı kullanılarak yapılmaktadır.

2 boyutlu koordinat düzlemi üzerine rastgele bir şekilde dağılmış noktalar olduğunu düşünelim. Birinci temel bileşenin bulunabilmesi için öncelikle koordinat sistemi üzerinde (0,0) noktası üzerine bir eğri rastgele bir şekilde yerleştirilir [15].

Burada her bir noktanın orijin üzerinde rastgele oluşturulan eğriye olan dik uzaklığına c dersek, bu c değeri pisagor teoremi kullanılarak elde edilmiş olur. C değerleri bütün noktalar için hesaplandığında her bir c değerinin orijine olan uzaklığının karesinin toplamının en yüksek olacak şekilde elde edilmesini sağlayan eğriye birinci temel bileşen eğrisi denir. Birinci temel bileşen eğrisi için yapılan işlem birinci temel bileşene ve orijine göre ikinci bir temel bileşene uygulandığında oluşan iki temel bileşen yeni x ve y koordinat sistemini oluşturmaktadır.

Veri seti üzerine doğrusal bir regresyon uygulanamayacağı test edildikten sonra doğrusal olmayan regresyon modelleri test edilmiştir.

Tablo 5.1. Doğrusal olmayan regresyon modeli uygulanan yöntemlerin sonuçları

		Doğruluk Oranı	Ortalama hata
Normal Veri	DecisionTreeRegressor	0.99944	0.09681
Normal Veri	BaggingRegressor	0.81631	5.67910
Normal Veri	RandomForestRegressor	0.71741	6.39991
PCA Uygulanmış Veri	DecisionTreeRegressor	1.0	0.0
PCA Uygulanmış Veri	BaggingRegressor	0.83643	6.25376
PCA Uygulanmış Veri	RandomForestRegressor	0.75573	7.07028

Doğrusal olmayan modeller veri seti üzerinde uygulandığında doğrusal modellere göre çok daha iyi sonuçlar verdiği görülmüştür. Bunun yanı sıra normal veri ile pca uygulanmış veri arasında doğrusal olmayan modeller açısından bir karşılaştırma yapılacak olur ise pca edilmiş verinin normal veriye göre daha yüksek bir doğruluk oranıyla çalıştığı ilgilendiğimiz doğrusal olmayan regresyon modelleri üzerinde gözlemlenmiştir.

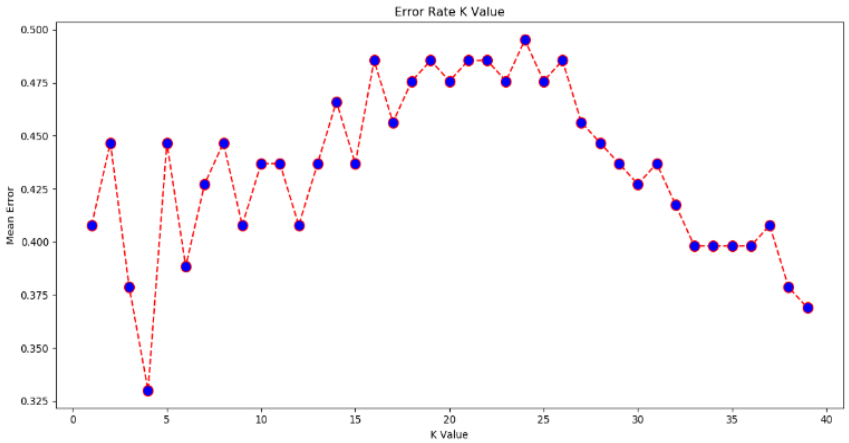
5.3. Sınıflandırma

Orman yangınları veri setinin sınıflandırma bölümünde k en yakın komşu algoritması kullanarak sınıflandırma ile lojistik regresyon kullanarak sınıflandırma işlemlerinin karşılaştırılması ve bu sınıflandırma modelleri üzerinde kullanılan veri setlerinin karşılaştırılması için ise normal veri setinden gelen değerler ile temel bileşen analizi yapılmış veri setinden gelen veriler kullanılarak karşılaştırma yapılmıştır.

Bilindiği üzere orman yangınları veri seti içerisinde bulunan çıktı kolonu içerisinde bulunan değerler sürekli değerlerden oluşmaktaydı. Sınıflandırma işlemi kategorik veriler (Orman yangını var / yok) arasında yapılabileceği için çıktı kolonu içerisinde bulunan sürekli değerleri ikili(binary) değerlere çevirilmesi gerekmektedir. Bu ikili değerlere çevirme işlemi orman yangını olmayan yerlere ‘0’ ve herhangi bir şekilde az ya da çok orman yangını

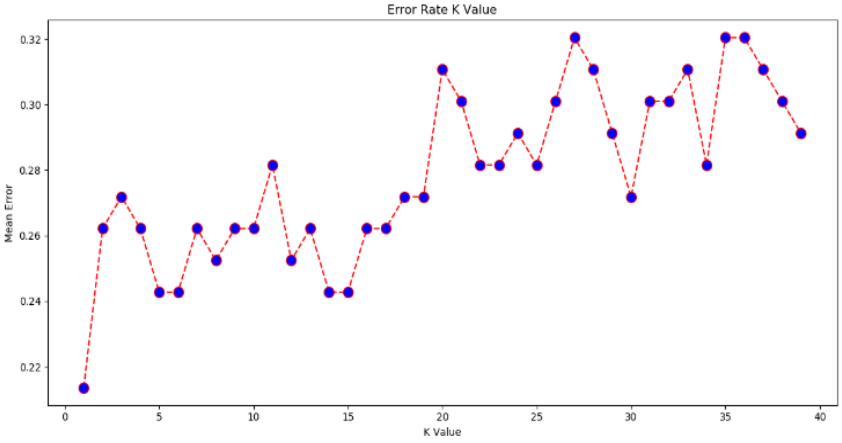
yaşanmış yerlere ‘1’ atanarak işlem yapılmıştır. Fakat burada çok yüksek yanmış alan değerine sahip olan değerler sınıflandırmanın doğruluk oranının arttırılması için veri seti içerisinden çıkarılmıştır.

K en yakın komşu sınıflandırma algoritması kullanılırken yapılan sınıflandırmada 1 ile 40 en yakın komşu algoritması kullanılarak yapılan veri eğitim işleminde 4 en yakın komşu değeri kullanıldığında en yüksek doğruluk oranına erişilmiştir.



Şekil 5.8. Normal veri için k en yakın komşu değerlerine göre ortalama hata grafiği

Yukarıdaki grafikte bulunan k en yakın komşu ve ortalama hata oranı göz önünde bulundurulduğunda, 4 komşu kullanılarak en az hata oranında işlem yapıldığı görülmektedir. Tablo 5.2. de bulunan knn algoritması için normal veri üzerinde 4 komşu kullanılarak yapılan eğitim sonucunda alınan doğruluk oranları baz alınmıştır.



Şekil 5.9. PCA uygulanmış veri için k en yakın komşu değerlerine göre ortalama hata grafiği

Yukarıdaki grafikte temel bileşenler analizi uygulanmış veri seti üzerinde knn algoritması çalıştırıldığında 1 en yakın komşu değeri kullanılarak en az hata alındığı görülmektedir. Tablo 5.2. de knn algoritmasının pca verisi için kullanıldığı bölümde alınan doğruluk oranı 1 en yakın komşu için geçerlidir ve en yüksek doğruluk oranı da bu komşu değerinde alınmıştır.

Tablo 5.2. Sınıflandırma modelleri uygulanan yöntemlerin sonuç tablosu

	Eğitim verisi doğruluk oranı	Test verisi doğruluk oranı	Çapraz Doğrulama Sonuçları
KNN Normal Veri	0.72	0.67	0.505
KNN PCA Veri	0.86	0.74	0.679
Lojistik Regresyon Normal Veri	0.58	0.52	0.40
Lojistik Regresyon PCA Veri	0.77	0.69	0.75

Tablo 5.2. de yapılan karşılaştırmada görüldüğü üzere kullanılan iki sınıflandırma algoritması da baz alındığında temel bileşen analizi yapılmış veri üzerindeki sınıflandırma işleminde çok daha iyi sonuçlar elde edilmiştir. Bunun yanı sıra KNN algoritması hem normal veri hemde temel bileşen analizi yapılmış veri üzerinde lojistik regresyondan daha iyi çalışmaktadır.

6. SONUÇLAR

- 517 satır 13 sütun olan orman yangınları veri seti içerisinde bulunan nümerik olmayan değerler nümerik değere çevirilerek temel bileşenler analizine hazır hale getirilmiştir.
- Temel bileşenler analizi sonucu veri seti içerisinde bulunan 13 özellik, birbirinden bağımsız ve korelasyonsuz 3 temel bileşene düşürülmüştür.
- Temel bileşeneler analizi yapılmış veri seti ve orijinal veri seti üzerinde doğrusal ve doğrusal olmayan regresyon analizi teknikleri test edilmiştir. Orman yangınları veri seti yapısı bakımından çok zor bir regresyon problemidir. Doğrusal regresyon modellerinin bu veri seti üzerine uygulanamadığı test edilmiştir.
- Doğrusal olmayan regresyon modelleri, orman yangınları veri seti için doğrusal regresyon analizlerinden çok daha iyi sonuç vermektedir.
- Orman yangınları veri seti üzerine uygulanan sınıflandırma metotları göz önüne alındığında hem orijinal veri seti hem de temel bileşen analizi uygulanmış veri seti göz önünde bulundurulduğunda k en yakın komşu algoritması, lojistik regresyondan daha iyi sonuç vermektedir.
- Temel bileşen analizi uygulanmış veri seti, orijinal veri seti ile karşılaştırıldığında sınıflandırma teknikleri bakımından orijinal veri setinden daha iyi sonuç vermektedir.
- Genel olarak bu çalışmada veri temizleme işlemleri, temel bileşenler analizi ile boyut indirgeme işlemleri, orijinal veri ile temel bileşenler analizinden elde edilen veri üzerine doğrusal ve doğrusal olmayan regresyon tekniklerinin uygulanıp uygulanamadığı aynı zamanda bu iki ayrı veri setinin doğrusal olmayan regresyonda kullanılan tekniklere göre başarımlarının karşılaştırılması, yine iki veri setinin sınıflandırma tekniklerinde çeşitli modellere göre başarımlarının karşılaştırılması bu çalışma içerisinde incelenmiştir.

7. KAYNAKÇA

[1] Capital. [Çevrimiçi] Eylül 1, 2009. https://www.capital.com.tr/kose-yazisi/kose-yazisi-433671/orman-yanginlaribilinclendirmenin-onemi.

[2] P. CORTEZ and A. MORAIS. A Data Mining Approach to Predict Forest Fires using Meteorological Data. In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimarães, Portugal, pp. 512-523, 2007.

[3] ALKAN, Ömer. Temel Bileşenler Analizi ve Bir Uygulama Örneği. Atatürk Üniversitesi Sosyal Bilimler Enstitüsü İşletme Anabilim Dalı, Erzurum 2008.

[4] GÜLTEKİN, Fikret. [Çevrimiçi] http://w3.balikesir.edu.tr/~bsentuna/wp-content/uploads/2013/03/Regresyon-Analizi.pdf.

[5] ÜNLÜ, Mustafa. Özdeğerler ve Özvektörler. [Çevrimiçi]

[6] GÜLTEKİN, Fikret. [Çevrimiçi] http://w3.balikesir.edu.tr/~bsentuna/wp-content/uploads/2013/03/Regresyon-Analizi.pdf.

[7] David M DIEZ, Christopher D BARR, Mine CETİNKAYA RUNDEL. OpenIntro Statistics, Third Edition.

[8] SERİN, Tarhan. 2010. Doğrusal Olmayan Regresyon Modellerinde Parametre Tahmin Yöntemleri, Öneriler ve Karşılaştırmaları.

[9] ORMAN, Mehmet ve GÜRCAN, Safa. Doğrusal olmayan regresyon analizi ve biyoistatistikte kullanımı. Ankara Üniversitesi, Veteriner Fakültesi, Biyometri Anabilim Dalı, Ankara 2001.

[10] Analytics Vidhya [Çevrimiçi] https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

[11] Stack Abuse [Çevrimiçi] https://stackabuse.com/k-nearest-neighbors-algorithm-in-python-and-scikit-learn/

[12] BAŞ, Metin ve ÇAKMAK, Zeki. Gri İlişkisel Analiz ve Lojistik Regresyon Analizi ile İşletmelerde Finansal Başarısızlığın Belirlenmesi ve Bir Uygulama. Anadolu Üniversitesi Sosyal Bilimler Dergisi.

[13] ÇAKIR, Mehmet ve TOPUZ, Derviş. Lojistik Regresyon Analiz Tekniğinin Eğitim Bilimi Araştırmalarında Uygulanabilirliği İle İlgili Bir Araştırma. Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi.

[14] Natural Resources Canada. [Çevrimiçi] http://cwfis.cfs.nrcan.gc.ca/background/summary/fwi.

[15] Josh STARMER [Çevrimiçi] https://statquest.org/video-index/