T.C.
SAKARYA ÜNİVERSİTESİ
BİLGİSAYAR VE BİLİŞİM BİLİMLERİ FAKÜLTESİ

BSM 401 BİLGİSAYAR MÜHENDİSLİĞİ TASARIMI

LİNEER REGRESYON PROGRAMI

G130910051 -M. Mert SEMİZ
G131210037 – Eyüp ESİN

Bölüm : BİLGİSAYAR MÜHENDİSLİĞİ
Danışman : Arş.Gör.Dr. MUSTAFA AKPINAR

2017-2018Güz Dönemi

ÖNSÖZ

Günümüzde birçok işletme rekabet üstünlüğünü elde etmede Big Data'nın önemini anlamış ve müşteri memnuniyetinin arttırılması, müşteri kazanma, hitap edebileceği potansiyel müşterileri bulma gibi konularda son derece başarılı teknikler olduğunu görmüştür.Bu projemize bu tekniklerden birinin program haline getirilmesi amacıyla yapılmıştır.

Bu çalışmamızın hazırlanması sürecinde, bize yardımcı olan, yol gösteren, emeği geçen hocamız MustafaAkpınar'a, bu zamana kadar bize maddi manevi destek olan ailemize teşekkürü borç biliriz.

İÇİNDEKİLER

SİMGELER VE KISALTMALAR LİSTESİ

MSE	: Mean squared error
SSE	: Sum of squared errors
MAPE	: Mean absolute percentage error
T	: Tahmin edilen y değeri
G	: Gerçek y değeri

ŞEKİLLER LİSTESİ

TABLOLAR LİSTESİ

ÖZET

Anahtar kelimeler: Big Data, Lineer Regresyon, Veri Madenciliği

Zamanımızda internetin çok fazla kişi tarafından kullanılması ve bununla beraber günlük hayatımızda aldığımız hizmetlere de internet üzerinde bulunan birçok uygulama yazılımları ile çok daha kolay çok daha hızlı erişebiliyor olmamız sebebiyle ortaya çıkan yaygın kullanım müşterilerin memnuniyetini sağlamak, hedef kitleye nasıl ulaşılabileceğini öğrenmek vb. birçok sebep ile son kullanıcının her türlü bilgisinin alınıp saklanmasına neden oldu.Bu karmaşık ve kalabalık bilgiler kullanılarak mantıklı veriler ortaya çıkarmanın birçok yolu ortaya çıkmıştır.Bu çalışma ilebu tekniklerden biri olan lineer regresyon denkleminin program ortamında gerçekleştirilmesi yapıldı.

BÖLÜM 1. GİRİŞ

1.1. Amaç

Bu çalışma ile büyük veya küçük boyutlu veriler kullanılarak, veri madenciliği yöntemlerinin en başında gelen ve en çok kullanılan yöntem olan lineer regresyon denkleminin program ortamında gerçekleştirilmesi yapılmıştır. Gerçekleştirdiğimiz proje ile bu büyük veya küçük verilerden anlamlı bilgiler çıkarılması vebu gerçekleştirme sonucunda bazı hata değerleri bulunarak hem yazılı hem grafiksel olarak gösterilmesi amaçlanmıştır.

1.2. Big Data

Son yıllarda, önemi gittikçe artan "Big Data", adı üstünde "büyük" beklentilere de sebep olmakta.Özellikle IoT kavramının hayatımıza girişi ile her nesnenin internete bağlanarak akıl kazanması, bunun yanında dijital uygulamalar ve servislerdeki sürekli artış ile birlikte, veriler çok çeşitli kaynaklardan şaşırtıcı bir hızla toplanmakta. Tüketicilerin haberi olmadan datanın toplanması, işlenmesi ve tekrar müşteriler ile iletişime geçmek için kullanılması, ülkemiz dahil pek çok yerde ciddi tartışma konusu olsa da; gerçek şu ki, bugün pek çok şirket müşterilerinin neyi, nereden nasıl aldığını; bağlı oldukları sosyal grupları ya da üye oldukları dernekleri; nerede ne kadar süre vakit geçirdiklerini ve hobilerini içeren muazzam bir dataya sahipler. Bu nedenle, doğru kullanıldığı ve değerlendirildiği vakit datanın altın kadar değeri var.

Her şirket, sürekli olarak ve artarak müşteri datalarını toplamaya devam ederken, bunların çok azı, toplanan datayı, müşteri ilişkilerini iyileştirmede ve müşteri memnuniyeti yaratmada kullanabiliyor. Aslında, datadan para kazanmak, karlı iş modelleri geliştirmek oldukça zor. Genellikle işletmeler ya da start-up lar "önce datayı toplayayım, belirli bir büyüklüğe geldikten sonra ne yapacağımıza bakarız ya

da en kötü toplanan datayı, bununla ilgilenen bir şirkete satarız" düşüncesiyle kolaya kaçmaktalar. Baştan belli bir amaç doğrultusunda toplanmayan datanın değerlendirilmesi ne yazık ki düşünüldüğü gibi kolay olmuyor ve yapılan çalışmaların karşılığı alınamıyor.Bu nedenle pek çok data odaklı şirketin başarısız olduğunu gözlemlemekteyiz.Gerçek şu ki data tek başına bir değer ifade etmiyor; "değer" bir ihtiyacı ya da problemi çözecek şekilde datanın işlenmesi ile oluşuyor.Bu nedenle de datayı anlayarak, eldeki datadan doğru çıktılar sağlamak çok daha fazla önem arz ediyor. [1]

1.2.1. Big Data Bileşenleri

Big Data (Büyük Veri)'nin oluşumunda genel literatüre göre 5 bileşen vardır. Bu bileşenler sırasıyla; "variety", "velocity", "volume", "verification" ve "value" olarak sıralanır. Genel olarak 5v şeklinde adlandırılmaktadır.

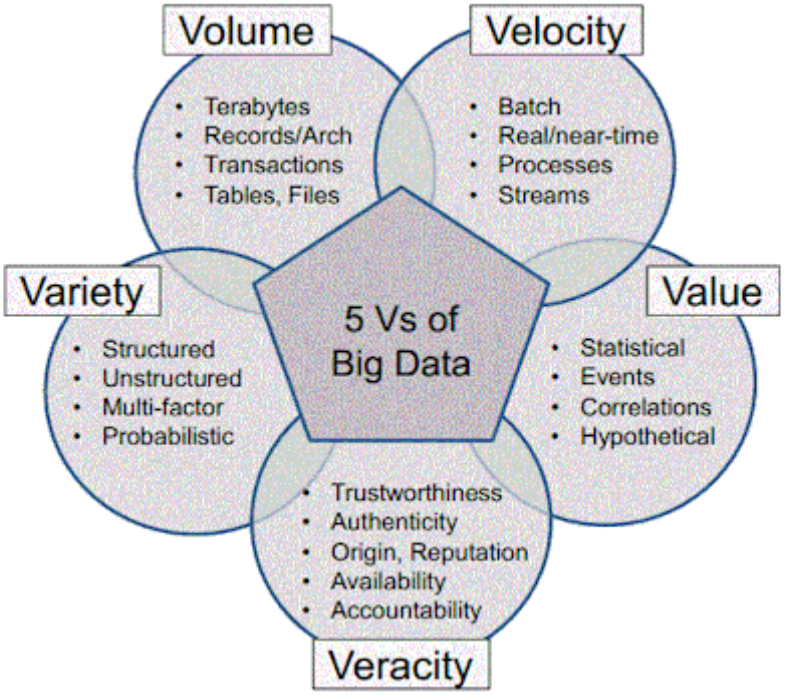
Variety (Çeşitlilik): Üretilen veriler genel olarak yapısal olmadığı ve bir çok farklı ortamdan elde edilen veri formatlarından oluştuklarından dolayı bütünleşik ve birbirlerine dönüştürülebiliyor olmaları gerekmektedir.

Velocity (Hız): Big data üretimi her geçen gün hızına hız katmakta ve bu veriler saniyede inanılmaz boyutlara ulaşmaktadır. Hızlı büyüyen veri, o veriye muhtaç olan işlem sayısının ve çeşitliliğinin de aynı hızda artması sonucunu ortaya çıkartmaktadır ve hem yazılımsal hemde donanımsal olarak bu yoğunluğu kaldırabilmeliyiz.

Volume (Veri Büyüklüğü): Big Data olarak isimlendirdiğimiz verilerimiz her geçen gün hızına hız katarak artıyor olabilir, haliyle gelecekteki durumlarımızı da ön plana koyarak ileride bu veri yığınları ile nasıl başa çıkacağımızı iyi düşünmemiz ve planlarımızı bu doğrultuda yapmamız gerekmektedir.

Verification (Doğrulama): Bu kadar hızlı büyüyen verilerin akışı sırasında gelen verilerin güvenli olup olmadığını kontrol etmemiz gerektiği durumlarda da bir diğer veri bileşeni olarak Verification (Doğrulama) görülebilir. Bu veri doğru kişiler tarafından görülebilir veya saklı kalması gerekiyor olabilir.

Value (Değer): Belkide en önemli katmanlardan bir tanesi de “Değer” katmanıdır, verilerimiz yukarıdaki veri bileşenlerinden filtrelendikten sonra büyük verinin üretimi ve işlenmesi katmanlarında elde edilen verilerin şirketimiz için artı değer sağlıyor olması gerekiyor.



Şekil 1.1. Big Data Bileşenleri[2]

1.3. Veri Madenciliği ve Teknikleri

Veri madenciliği, büyük ölçekli veriler arasından bilgiye ulaşma, bilgiyi madenleme işidir.Ya da bir anlamda büyük veri yığınları içerisinden gelecek ile ilgili tahminde bulunabilmemizi sağlayabilecek bağıntıların bilgisayar programı kullanarak aranmasıdır.

Veri madenciliğinin modelleri ve bu modellerin bir çok tekniği vardır. Bu Modelleri ve tekniklerini kısaca şöyle sıralayabiliriz “Karar ağaçları, yapay sinir ağları, Naïve Bayes, bulanık mantık, bellek temelli nedenleme, k-means, lineer regresyon” veri

madenciliği yöntemlerinin en başında gelen ve en çok kullanılanı lineer regresyonlardır.

BÖLÜM 2. YÖNTEM

2.1. Regresyon Analizi ve Önemi

İstatistiğin en sık kullanılan yöntemlerinden birisi hiç kuşku yok ki Regresyon Analizidir.Çoğu istatistiksel yöntemlerin temeli sayılabilecek ve araştırmalarda sık sık başvurulabilecek yöntemlerin başında gelir.Regresyon, bir ya da birden çok değişkeni bir ya da birden çok bağımsız olarak adlandırılan değişkenlere bağlama işi ve biçimidir. Regresyon analizinde çok seçenek vardır: değişkenlerin türlerine göre, dağılımlara göre, model varsayımlarına göre, verilerdeki bozulmalara göre, hatta gözlem ve değişken sayısına göre... Bir regresyon analizinin yapılması kısa süreli bir iş değildir; bir sürecin uygulanmasıdır. Ayrıca paket programların regresyon çıktıları da çoğu zaman mekaniktir; en iyi modeli ya da modelleri gösterme garantileri yoktur. O nedenle model kurucunun analiz sürecini iyi bilmesi ve sürecin aşamalarını iyi uygulaması gereklidir. [3]

2.1.1. Regresyon Kullanım Alanları

Regresyon modelleri farklı amaçlarla kullanılmaktadır. Bunlar bazılar aşağıdaki gibidir:

1. Veri tanımlama Parametre kestirimi
3. Kestirim ve önkestirim
4. Denetleme

Mühendisler ve bilim adamları genellikle bir veri kümesini özetlemek ya da tanımlamak için denklemler kullanır.Regresyon analizi bu tür denklemlerin geliştirilmesine yardımıcı olmaktadır.Örneğin, önemli boyutta teslim süresi ve teslim hacmi verileri toplayabiliriz ve bir regresyon modeli, söz konusu veriye ilişkin tablo ya da grafiklerden daha uygun ve yararlı bir özet sunabilir.Bazı durumlarda parametre kestirim problemleri regresyon yöntemleriyle çözülebilmektedir. Örneğin, kimya mühendisleri tepkinin hızı y ve konsantrasyon x arasındaki ilişki

$$y = \beta_1 x/ (x + \beta_2)+ \varepsilon \quad \textbf{(2.1)}$$

Denklem (2.1) Michaelis-Menten denklemini tanımlamak için kullanılır.Bu modelde β_1 reaksiyonun asimtotik hızını, konsantrasyonun arttıkça elde edilen maksimum hızı göstermektedir. Eğer farklı konsantrasyon değerlerinde gözlenen hız değerlerine ilişkin bir örneklem varsa mühendis regresyon analizini bu modeli verilere uydur maksimum hızın bir kestirimini bulmak için kullanabilir.[3]

Regresyonun birçok kullanım alanı yanıt değişkeninin önkestirimini içermektedir.Örneğin, belli sayıdaki meşrubat için gereken teslim süresini kestirmek isteyebiliriz.Bu rota belirleme ve zamanlama gibi teslim etkinliklerini planlamada teslim işlemlerinin verimliliğini değerlendirmede yardımcı olabilir. [3]

Regresyon modelinin kullanımında model ya da hatasından kaynaklanabilecek diş değer bulma tehlikesi vardır. Ancak, model biçimi doğru olsa da model parametrelerinin zayıf kestirimi zayıf önkestirim performasına yol açabilir.Regresyon modelleri, denetleme amacıyla da kullanılabilir.Örneğin, bir kimya mühendisi, kâğıdın gerilme direncini hamurdaki ağaç konsantrasyonuyla ilişkilendiren bir model geliştirmek için regresyon analizini kullanmak isteyebilir.Bu denklem de ağaç konsantrasyon düzeyini değiştirerek gerilme direncini uygun değerler arasında tutabilmek için kullanılabilir. [3]

Regresyon denklemi denetleme amacıyla kullanıldığında değişkenlerin neden-sonuç ilişkisi içinde olmaları önemlidir.Denklem yalnızca önkestirim amaçlı kullanılıyorsa bir neden-sonuç ilişkisine gerek olmadığına dikkat ediniz.Bu durumda yalnızca regresyon denklemini oluşturmak için kullanılan verilerdeki ilişkilerin hâlâ geçerli olması yeterlidir.Örneğin, Atlanta Georgia ağustos ayındaki orijinal günlük elektrik tüketimi ağustos ayındaki en yüksek sıcaklık için iyi bir önkestirim modeli olabilir.Ancak, elektrik tüketimini azaltarak en yüksek sıcaklığı aşağı çekmek olanaksızolacaktır. [3]

2.2. Regresyon Türleri

2.2.1. Basit Lineer Regresyon

Basit Lineer Regresyonda, y yanıt değişkeni ile doğrusal ilişkiye sahip tek bir x bağımsız değişkeninin bulunduğu basit doğrusal regresyon modeli ele alinmaktadır. Bu basit doğrusal regresyon modeli aşağıdaki gibidir: [3]

$$y = \beta_0+\beta_1 x+ \varepsilon \quad \textbf{(2.2)}$$

Burada β_0 kesim noktası ve β_1 eğim olmak üzere bilinmeyen sabitlerdir ve ε da rastgele hata bileşenidir.Hatalarin sifir ortalamaya ve bilinmeyen varyansa sahip oldukları varsayılır.Buna ek olarak hataların ilişkisiz olduğunu varsayarız.Bu da, bir hatanın değerinin başka bir hatanın değerine bağlu olmaması anlamına gelmektedir.x değişkenini, veri analisti tarafından kontrol edilen ve göz ardı edilebilir hatayla ölçülen bir bağımsız değişken olarak, y yanıtını da bir rastlantı değişkeni olarak kabul etmek uygundur. Diğer bir deyişlex'in her olasi değerine karşılık y için bir olasılık dağılımı vardır. Dağılımın ortalamasi:

$$E(y|x) = \beta_0 + \beta_1 x \quad \textbf{(2.3)}$$

Denklem (3)'deki gibidir ve varyansıda aşşağıdaki gibidir:

$$Var(y|x) = Var(\beta_0 + \beta_1 x+ \varepsilon) = \sigma^2 \quad \textbf{(2.4)}$$

Böylece, y'nin varyansı x'in değerine bağlı olmasa da y'nin ortalaması x'in bir doğrusal fonksiyonudur.Dahası hatalar ilişkisiz olduğundan yanıtlar da ilişkisizdir.$\beta_0$ve β_1 parametreleri genellikle regresyon katsayıları olarak adlandırılır. Bu katsayıların basit ve genellikle yararlı yorumları vardır.β_1 eğimi, x'teki bir birim değişiklik ile elde edilen y'nin dağılımının ortalamasındaki değişikliktir. Eğer x üzerindeki verilerin araliği x = 0'ı içeriyorsa β_0 kesim noktası x = 0 olduğunda y yanıt dağılımının ortalamasını verir. Eğer x'in aralığı 0 değerini içermiyorsa β_0 'ın kullanışlı bir yorumlanması yoktur.[3]

2.2.2. Çoklu Lineer Regresyon

Data for Multiple Linear Regression

		Regressors			
Observation, i	Response, y	x_1	x_2	. . .	x_k
1	y_1	x_{11}	x_{12}	. . .	x_{1k}
2	y_2	x_{21}	x_{22}	. . .	x_{2k}
⋮	⋮	⋮	⋮		⋮
n	y_n	x_{n1}	x_{n2}	. . .	x_{nk}

Şekil 2.1. Çoklu Lineer Regresyon Dataları[3]

Bu bölüm boyunca $x_1,x_2,\ldots,x$bağımsız değişkenlerinin, sabit (başka bir deyişle, matematiksel ya da rastgele olmayan) değişkenler oldukları ve hatasız ölçüldükleri varsayılacaktır. Bununla birlikte basit doğrusal regresyonda tartışılan sonuçlar, bağımsız değişkenlerin rastlanti değişkeni olduğu tüm durumlar için de geçerlidir.Bu, oldukça önemlidir; regresyon verileri bir gözlemsel çalişmadan elde edildiği zaman bağımsız değişkenlerin bir ya da birden çoğu rastlantı değişkeni olabilmektedir.Veriler bir deney tasarımından elde edildiğinde x'lerin sabit değişkenler olma olasılığı daha fazladır.x'ler rastlantı değişkenleri olduğu zaman her bir bağımsız değişkendeki gözlemler bağımsız olmalı ve dağılım, regresyon katsayılarına β 'lar ya da σ^2'ye bağlı olmamalıdır. Hipotezler test edilirken ya da güven aralıkları oluşturulurken x_1, x_2, ...,x_k verildiğinde y'nin koşullu dağılımının $\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_kx_k$ olduğunu varsayacağız. Örneklem regresyon modeli aşağıdaki gibi yazılabilir:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2+\ldots} \beta_k x_{ik} + \varepsilon_i$$

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i \; i = 1,2 \ldots, n \qquad \textbf{(2.5)}$$

En küçük kareler fonksiyonu ise aşşağıdaki gibidir:

$$s(\beta_0, \beta_1, \ldots, \beta_k = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_y)^2 \qquad \textbf{(2.6)}$$

S fonksiyonu, $\beta_0,\beta_1,\beta_{2,\ldots,}$ β_k'ya gore minimize edilmelidir. $\beta_0,\beta_1,\beta_{2,\ldots,}$ β_k'nın en küçük kareler kestiricileri,

$$\frac{\partial S}{\partial \beta_0}|_{(\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_k)} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 \sum_{j=1}^{k}\beta_j x_{ij}) = 0$$

$$\frac{\partial S}{\partial \beta_0}|_{(\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_k)} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 \sum_{j=1}^{k}\beta_j x_{ij})x_{ij} = 0 \quad j = 1,2,\ldots,k \qquad \textbf{(2.7)}$$

eşitliklerini sağlamalıdır. Denklem (2.7)'nin sadeleştirilmesiyle en küçük kareler normal denklemleri elde edilir. [3]

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \cdots + \beta_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}\,x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{i1}\,x_{ik} = \sum_{i=1}^{n} x_{i1} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik}x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik}\,x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 = \sum_{i=1}^{n} x_{ik} y_i \qquad \textbf{(2.8)}$$

Bilinmeyen regreyon katsayılarının her biri için bir denklem olmak üzere p = k + 1 sayıda denklem vardır. Normal denklemlerin çözümü, $\beta_0,\beta_1,\beta_{2,\ldots,}\ \beta_k$ en küçük kareler kesicilerini verecektir. [3]

Çoklu regresyon modelleriyle ilgilenirken onları matris biçiinde ifade etmek daha uygundur. Bu, modelin, verilerin ve sonuçların daha kısa ve öz bir şekilde gösterilmesini sağlar. [3]

Denklemin (2.5)'te verilen modelin matris gösterimi

$$y = X\beta + \varepsilon \qquad \textbf{(2.9)}$$

şeklindedir ve burada matrisler aşşağıdaki gibidir:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, (15) \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

$$\beta=\begin{bmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_k\end{bmatrix}, (17) \quad \varepsilon=\begin{bmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{bmatrix} \tag{2.10}$$

Genelde y, n x 1'lik gözlemler vektörü ; X, n x p'lik bağımsız değişkenler matrisi; β, p x 1'lik regresyon katsayıları vekttörü ve ε, n x 1'lk rastgele hatalar vektörüdür. [3]

Aşağıda verilen S(β) fonksiyonunu minimize ederek $\hat{\beta}$ en küçük kareler kestiriciler vektörünü elde ederiz[3]

$$S(\beta)=\sum_{i=1}^{n}\varepsilon_i^2=\varepsilon'\varepsilon=(y-X\beta)'(y-X\beta) \tag{2.11}$$

S(β) aşağıdaki gibi ifade edilebilir:

S(β)=y'y-β'X'y-y'Xβ+β'X'Xβ

$$=y'y-2\beta'X'y+\beta'X'X\beta \tag{2.12}$$

B'X'y, 1 x 1'lik bir matris ya da skaler olduğundan transpozu da (β'X'y)' = y'X β aynı şekilde skalerdir. En küçük kareler kestiricileri,

$$\frac{\partial S}{\partial \beta_0}|_{(\hat{\beta})}=-2X'y+2X'X\hat{\beta}=0 \tag{2.13}$$

eşitliğini sağlar. basitleştirilirse

$$X'X\hat{\beta}=X'y \tag{2.14}$$

Olur.Denklem (2.14) en küçük kareler normal denklemleridir. Bu denklemler

$$\frac{\partial S}{\partial \beta_0}|_{(\hat{\beta})}=-2X'y+2X'X\hat{\beta}=0 \tag{2.15}$$

bu skaler gösterimin matris karşılığıdır. Normal denklemleri çözmel için bu matris karşılığının her iki yanı X'X in tersi ile çarpılır böylece, β'nin en küçük kareler kestiricisi,

$$\beta=(X'X)^{-1}X'y \tag{2.16}$$

olur; burada, $(X'X)^{-1}$ ters matrisinin de mevcut olma koşulu sağlanmalıdır. Eğer bağımsız değişkenler doğrusal bağımsız ise X matrisinin hiçbir sütununun digger sütunlarla doğrusal bir birleşimi yoktur.(2.14)'deki normal denklemlerin matris formunun (2.8)'deki skaler form ile aynı olduğu kolayca görünür. (2.14) aşşağıdak gibi yazılırr;

$$\begin{bmatrix} n & \sum_{i=1}^{n}x_{i1} & \sum_{i=1}^{n}x_{i2} & \cdots & \sum_{i=1}^{n}x_{ik}\\ \sum_{i=1}^{n}x & \sum_{i=1}^{n}x_{i1}^2 & \sum_{i=1}^{n}x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n}x_{i1}x_{ik}\\ \vdots & \vdots & \vdots & & \vdots\\ \sum_{i=1}^{n}x_{ik} & \sum_{i=1}^{n}x_{ik}x_{i1} & \sum_{i=1}^{n}x_{ik}x_{i2} & \cdots & \sum_{i=1}^{n}x_{ik}^2 \end{bmatrix}\begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\\ \vdots\\ \hat{\beta}_k\end{bmatrix}=\begin{bmatrix}\sum_{i=1}^{n}y_i\\ \sum_{i=1}^{n}x_{i1}y_i\\ \vdots\\ \sum_{i=1}^{n}x_{ik}y_i\end{bmatrix} \tag{2.17}$$

Eğer yukarıdaki matris çarpımı gerçekleşirse normal denklemlerinin skaler dormu elde edilir.Bu gösterimde X'X, p x p simetrik matris ve X'y, p x 1 sütun vektörüdür. X'X matrisinin özel yapısını belirtelim: X'X in köşegen elemanlarii X'in Sütunlarındaki elemanların kareler toplamıdır; köşegen dışı elemanları ise X'in sütunlarındaki elemanların çapraz çarpımlarının toplamıdırç Ayrıca X'y, y_i gözlemleri ile Z sütunlarının çapraz çarpımlarının toplamıdır. [3]

Bağımsız değişkenlerin tüm x' = $[1,x_1,x_2,\ldots,x_k]$ düzeylerine karşılık gelen kestirim modeli şu şekilde gösterilir:

$$\hat{y} = x'\hat{\beta} = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_j \quad \textbf{(2.18)}$$

$y_i$gözlem değerlerine karşılık gelen $\hat{y}_I$ kestirim değerleri vektörü,

$$\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y = Hy \quad \textbf{(2.19)}$$

Olur. n x n boyutlu H = $X(X'X)^{-1}X'$ matrisi genellikle şağka matrisi olarak adlandırılır. Bu matris, gözlenmiş değerler vektörünü $\hat{y}_i$ kestirim değerleri vektörüyle eşleştirir.Şapka matrisi ve özellikleri regresyon analizinde merkezi bir rol oynamaktadır.$y_i$gözlem değerleri ile bunlara karşılık gelen $\hat{y}_I$ kestirim değerleri arasındaki $e_i = y_i - \hat{y}_I$ farkı artıklardır. n sayıdaki artıklar, matris biçiminde aşağıdaki gibi yazılabilir:

$$e = y - \hat{y} \quad \textbf{(2.20)}$$

Artıklar vektörü e, aşağıda gösterildiği gibi birçok farklı biçimde ifade edilebilir: [3]

$$e = y - X\hat{\beta} = y - Hy = (I - H)y[3] \quad \textbf{(2.21)}$$

2.2.3. Polinomiyal Regresyon

Y = $x\beta_0 + \varepsilon$ doğrusal regresyon modeli, β bilinmeyen parametrelerine göre doğrusal olan bir bağıntıyı uydurmada genel bir modeldir. Bu model, polinomiyal regresyon modellerinin önemli bir sınıfını da içerir. Örneğin, tek değişkenli ikinci dereceden bir polinomiyal model,

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x^2 + \varepsilon \quad \textbf{(2.22)}$$

ve iki değişkenli ikinci dereceden bir polinomiyal model,

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{12} x_1 x_2 + \varepsilon \quad \textbf{(2.23)}$$

biçiminde verilen doğrusal regresyon modelleridir.

Polinomiyal modeller, yanıt değişkeninin eğrisel olduğu durumlarda yaygın kullanılır hatta karmaşık doğrusal olmayan ilişkiler, x'lerin küçük aralıkları üzerinden polinomlar ile uygun biçimde modellenebilir. Bu bölümde polinomları uydurma ile ilgili çeşitli sorunlara ve konulara değinilecektir.[3]

BÖLÜM 3. MODELLEME

3.1. Yapılan İşlemler

Projede öncelikle veri setlerinin kayıtlı olduğu Excel seçimi ve Excel sayfasını seçme özelliğini VerileriGetir.cs sayfasında tasarladık.Excelden çekilen bu verileri dataGridView1 nesnesinde görüntüledik.Daha sonra Bağımlı (Y) ve Bağımsız (X) değişkenlerinin seçimlerini checkedListBox nesnesi yardımıyla Ayirim.cs sayfasında yazdık.

Amacımız $[Y]_{nx1}=[X]_{nxm}[C]_{mx1}$ formülündeki C matrisini (katsayıları) bulmak olduğu için islemler.cs sayfasında yazdığımız $(X'X)^{-1}X'Y=C$ formülünden yararlandık. Bu formülde öncelikle (X'X) bu işlemini yaptıktan sonra X'in transpozunu aldık ve X ile çarptık daha sonra (X'X) işleminin tersini almadan determinantının sıfır olup olmadığını HataKontrol.cs'de kontrol ettirdik eğer sonuç sıfir ise matrisin tersi alınamadığı için hata mesajı vermesini sağladık, daha sonra determinant sıfir dan farklıysa (X'X) işleminin tersinin alınmasını sağladık ve X'in transpozuyla çarptık. Son olarak çıkan sonucu Y matrisiyle çarparak katsayıları bulduk.C matrisini dataGridView2 nesnesinde değerleriyle birlikte gösterttik.Bulunan C değerlerine karşılık tahmin değerleri (Tahmin edilen Y değerleri) bulduk ve bu değerleri gerçek değerlerle birlikte dataGridView3 nesnesinde ve grafikte gösterttik.
Son olarak MSE,MAPE,R2,SSE hata değerlerini hesaplayıp dataGridView4 de görüntületirdik.

3.2. Hata Değerleri

T: Tahmin edilen y değeri
G: Gerçek y değeri
K: Başlangıç değeri
N: Bitiş değeri

$$\mathrm{MSE} = \frac{1}{n}\sum_{k=1}^{n}(T-G)^2 \quad \textbf{(3.1)}$$

Bir istatistiksel öğrenme metodunun belirli bir veri seti üzerindeki performansını değerlendirmek için metodun ürettiği tahminlerin gerçek sonuçlarla ne kadar örtüştüğünü ölçümlememize yarayacak yöntemlere ihtiyacımız var. Yani, belirli bir gözlem için tahmin edilen cevap(reponse) değerin o gözlemin gerçek cevap(reponse) değerine ne kadar yakın olduğunu sayısallaştırmamız gerekiyor. Regresyon için, en yaygın olarak kullanılan ölçüm Ortalama Karesel Hata (Mean Squared Error-MSE)'dır ve şu şekilde ifade edilir:

$$\mathrm{SSE} = \sum_{k=1}^{n}(T-G)^2 \quad \textbf{(3.2)}$$

Açıklanan kareler toplamı SSE'dir.

$$\mathrm{MAPE} = \frac{1}{n}\sum_{k=1}^{n}\left(100\,\frac{T-G}{G}\right) \quad \textbf{(3.3)}$$

Ortalama mutlak hata yüzdesi olarak bilinir ve başlıca hata ölçme yöntemlerinden birisidir.bir zaman serisindeki sapmaların gerçekleşen değerlere oranlarının ortalaması alınarak bulunur. Mümkün olduğunca sıfıra yakın olması istenir. Özellikle ortalamaya göre yüksek standart sapmaya sahip serilerde yüzde oranı olarak fikir vermesi nedeniyle mean absolute error a göre daha çok tercih edilir.

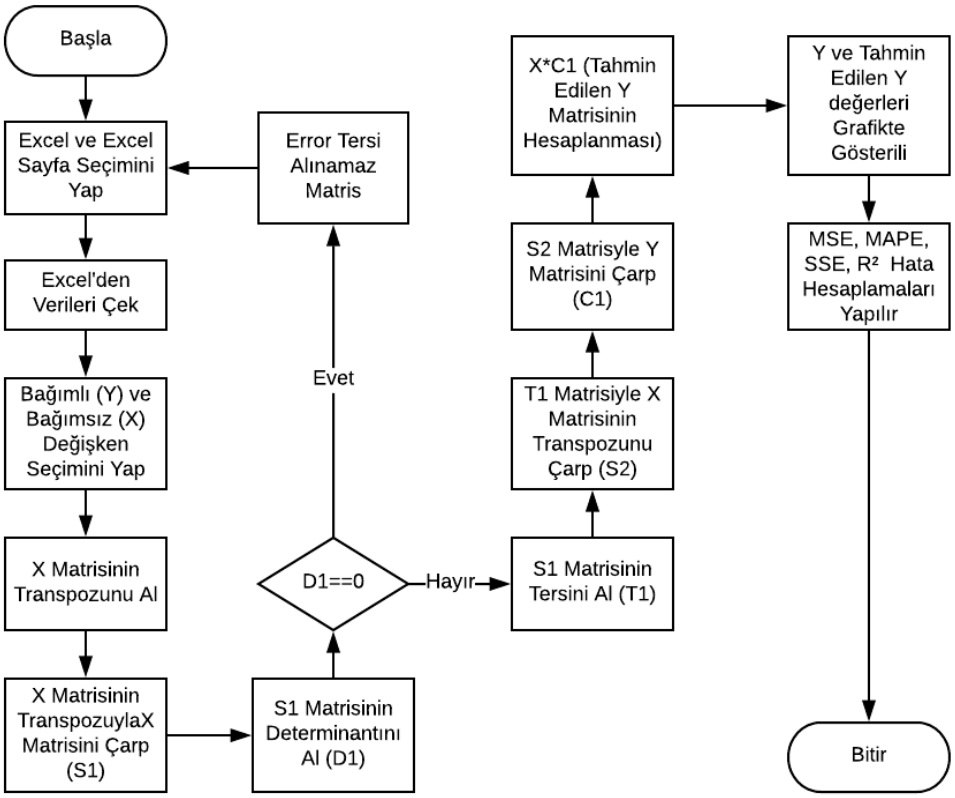
$$R^2 = \frac{\sum_{k=1}^{n}(T-\widehat{G})^2}{\sum_{k=1}^{n}(G-\widehat{G})^2} \quad \textbf{(3.4)}$$

Regresyon analizinde elde edilen denklemin bagimli degiskeni ölçme gücü.

3.3. Akış Diagramı

Programımızın çalışma şekli şu şekildedir; Önce Exel seçimi ve ardından excel dosyasının içinde bulunan sayfalar arasından seçim yapılır. Bu seçimin ardından excel verileri excelden çekilip programa aktarılır. Bağımsız ve bağımlı değişkenlerin seçimi kullanıcı tarafından yapıldıktan sonar program X matrisinin transpozunu alır, X matrsininin transpozuyla X matrsini birbirine çarpıp determinantını alır. Bu determinantın sıfıra eşit olup olmadığı kontrolü burda yapılır eğer sıfıra eşitse

programımız hata verir sıfıra eşit değilse X matrisi ile X matrsinin transpozunun çarpımının tersi alınır çıkan sonuç ile X matrisinin transpozu çarpılır ardından çıkan sonuç ile Y matrisi çarpılır. Bu sonuç ile X matrisi çarpılınca tahmin edilen Y matrisi hesaplanmış olur ve bunlar grafikte gösterilir.Ardında hata hesaplamalarımız yapılıp ekranda görüntülenir.



Şekil 3.5. Akış diyagramı

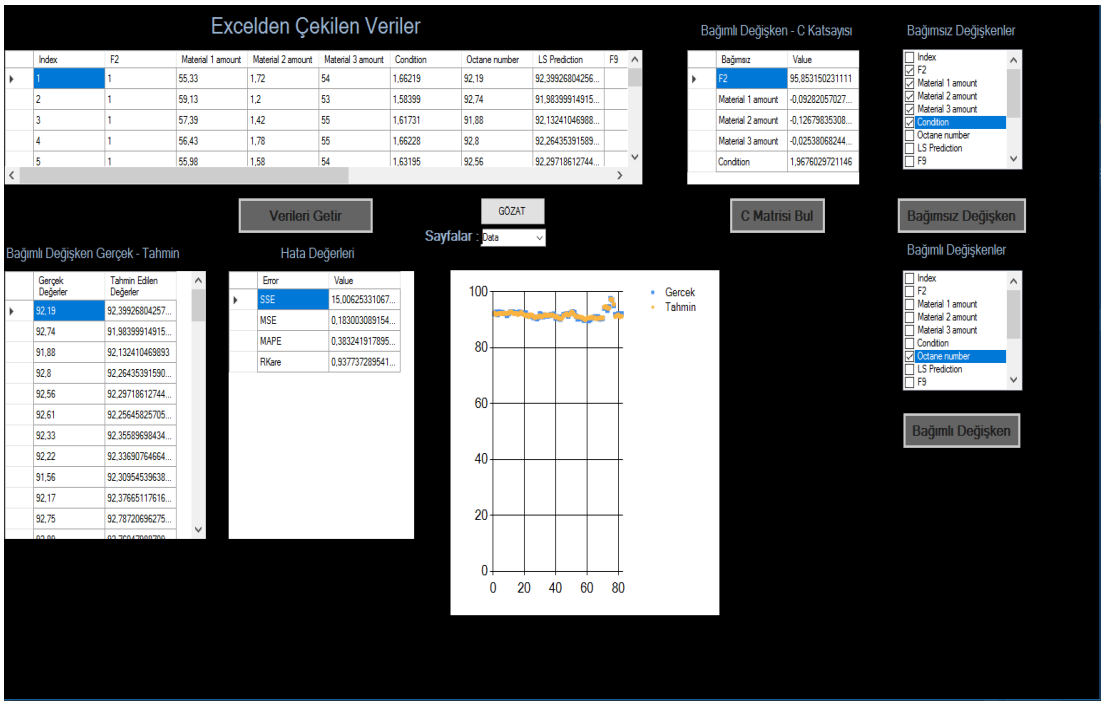
BÖLÜM 4. ÇALIŞMA SONUÇLARI

4.1. Yapılan Testler

Programımızı test ettiğimiz uci data setlerdinden[6] alınmış ana üç tane veri setine göz atalım.

4.1.1. Birinci Test

Birinci veri setinde hangi materyallerden kaç oktan kullanıldığını gösteren seksaniki veri vardır. Bu veri setinde programımız beş bağımsız bir bağımlı değişken seçilerek işleme sokuluyor. Bu işlemlerin ardından bulunan sonuçlarla bağımsız değişkenleri çarparak kendi buldumuz bağımsız değişkenleri ekrana yazdırıyoruz daha sonra "Gerçek" ve "Tahmin" edilen "Y" değerleri olarak grafik şeklinde incelenmek üzere ekranda gösterilir.



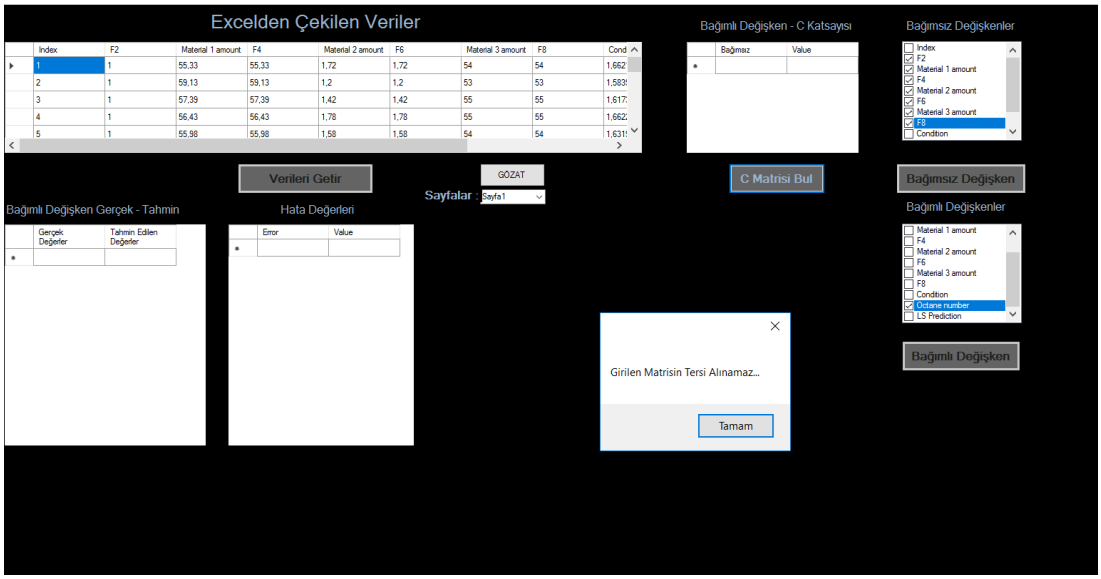
Şekil 4.1. 1. Datanın Program Çıktısı

4.1.2. İkinci Test

İkinci veri setinde veri sayısını üçyüz doksandokuza çıkararak arabaların bilgilerini içeren verilerle test ettik. Aynı işlemlerle test edilen bu veri setinde herhangi bir hata almadık.

4.1.3. Üçüncü Test

Üçüncü ve son veri setimizde yaptımız işlemlerden birinin yapılamaması durumunu incelicez buda matrisin tersini alma işlemimizde çok bariz ortaya çıkıyor. Matrisin tersinin alınabilmesi için determinantı sıfırdan farklı olmak zorunda bu durumda matrisin determinantı sıfıra eşit olursa programımız bunun özel bir durum olduğunu ve işlemi gerçekleştiremiyeceğini ifade eden bir mesajla bunu kullanıcıya söylemeli. Bu veri setinde determinant sıfır olan bir veri seti aldık ve test ettik ekran çıktısında görüldüğü üzere program bizi matrisin tersi alınamaz diyerek uyarı mesajı veriyor.



Şekil 4.2. 3. Datanın Program Çıktısı

BÖLÜM 5. SONUÇ VE ÖNERİLER

5.1. Sonuç

Big Datadan anlamlı veriler üretmek günümüzün çoğu alanının en önemli stratejisi haline gelmiştir.Big Datada ki karmaşık ve kalabalık bilgiler kullanılarak mantıklı veriler ortaya çıkarmanın birçok yolu ortaya çıkmıştır.Bu çalışma ile bu tekniklerden biri olan lineer regresyon denkleminin program ortamında gerçekleştirilmesi yapıldı.Ve ardımdam yapılan testler sonucunda programımızın girilen verileri doğru bir şekilde işleyip sonuçları bize doğru olarak verdiğini gördük.Hatalı durumlarda ise program hata verip hatanın sebebini söylemektedir.

5.2. Öneriler

İlerde programımızın çıkarttığı bu değerlere göre denklemin parametre testleri sonuçlara eklenecektir.Bunun dışında ise alt veri setleriyle (örneklem kümeleri) ile farklı sonuçlardaki parametrelerin davranışlarının görüneceği bir program olacaktır.

KAYNAKLAR

[1] http://digitalage.com.tr/2017-big-data-buyuk-veri-trendleri/

[2] https://tr.pinterest.com/pin/364932376032603552/

[3] Introduction to Linear Regression Analysis, Fifth Edition. Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining. © 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc

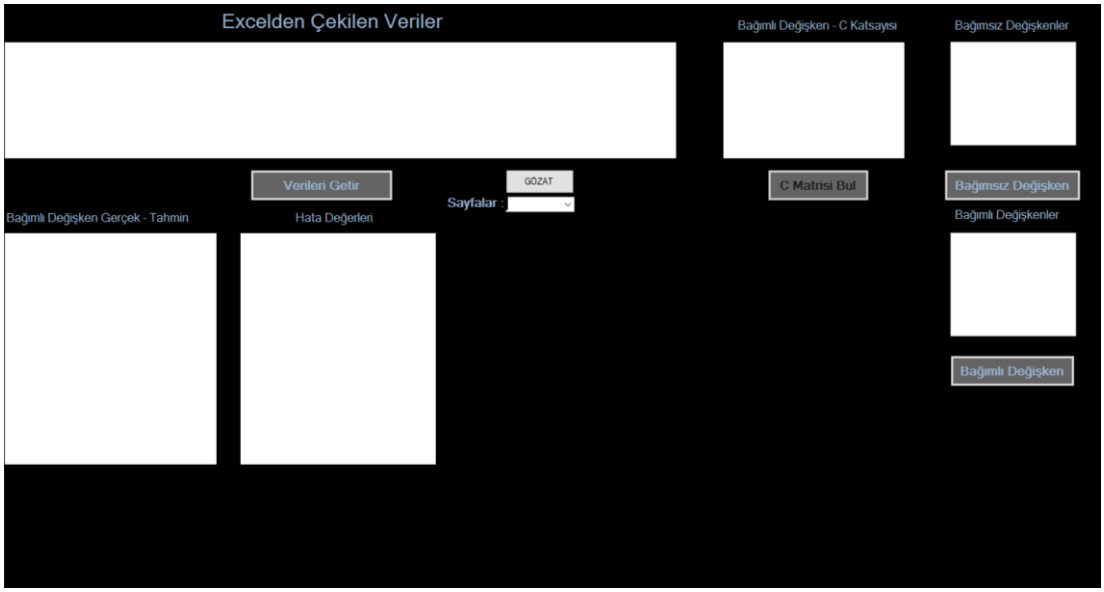
[4] https://archive.ics.uci.edu/ml/datasets.html

[5] http://www.google.com

[6] https://archive.ics.uci.edu/ml/datasets.html

EKLER

EK A: Programın Açılış Ekran Görüntüsü



ÖZGEÇMİŞLER

Eyüp Esin, 16.08.1995'te İstanbul'da doğdu. İlk, orta ve lise eğitimini İstanbul'da tamamladı. 2013 yılında Sakarya Üniversitesi Bilgisayar Mühendisliği Bölümü'nü kazandı. 2017 yılında Birtek Bilişim Sistemleri San. Tic. Ltd. şirketinde donanım stajını yapmıştır. Sakarya Üniversitesi Bilgisayar Mühendisliği Bölümünde eğitimine devam etmektedir.

M. Mert Semiz, 01.08.1994'te İstanbul'da doğdu. İlkokul ve ortaokul eğitimini Şenerbirsöz İlk öğretim okulunda ve lise eğitimini 2012'dePendik Asya Türktelekom Anadolu Teknik Lisesinde tamamladı. 2013 yılında Sakarya Üniversitesi Bilgisayar Mühendisliği Bölümü'nü kazandı. 2017 yılında Birtek Bilişim Sistemleri San. Tic. Ltd. şirketinde donanım stajını yapmıştır. Sakarya Üniversitesi Bilgisayar Mühendisliği Bölümünde eğitimine devam etmektedir.

BSM 401 BİLGİSAYAR MÜHENDİSLİĞİ TASARIMI DEĞERLENDİRME VE SÖZLÜ SINAV TUTANAĞI

KONU :
ÖĞRENCİLER (Öğrenci No/AD/SOYAD):

Değerlendirme Konusu	İstenenler	Not Aralığı	Not
Yazılı Çalışma			
Çalışma klavuza uygun olarak hazırlanmış mı?	x	0-5	
Teknik Yönden			
Problemin tanımı yapılmış mı?	x	0-5	
Geliştirilecek yazılımın/donanımın mimarisini içeren blok şeması (yazılımlar için veri akış şeması (dfd) da olabilir) çizilerek açıklanmış mı?			
Blok şemadaki birimler arasındaki bilgi akışına ait model/gösterim var mı?			
Yazılımın gereksinim listesi oluşturulmuş mu?			
Kullanılan/kullanılması düşünülen araçlar/teknolojiler anlatılmış mı?			
Donanımların programlanması/konfigürasyonu için yazılım gereksinimleri belirtilmiş mi?			
UML ile modelleme yapılmış mı?			
Veritabanları kullanılmış ise kavramsal model çıkarılmış mı? (Varlık ilişki modeli, noSQL kavramsal modelleri v.b.)			
Projeye yönelik iş-zaman çizelgesi çıkarılarak maliyet analizi yapılmış mı?			
Donanım bileşenlerinin maliyet analizi (prototip-adetli seri üretim vb.) çıkarılmış mı?			
Donanım için gerekli enerji analizi (minimum-uyku-aktif-maksimum) yapılmış mı?			
Grup çalışmalarında grup üyelerinin görev tanımları verilmiş mi (iş-zaman çizelgesinde belirtilebilir)?			
Sürüm denetim sistemi (Version Control System; Git, Subversion v.s.) kullanılmış mı?			
Sistemin genel testi için uygulanan metotlar ve iyileştirme süreçlerinin dökümü verilmiş mi?			
Yazılımın sızma testi yapılmış mı?			
Performans testi yapılmış mı?			
Tasarımın uygulamasında ortaya çıkan uyumsuzluklar ve aksaklıklar belirtilerek çözüm yöntemleri tartışılmış mı?			
Yapılan işlerin zorluk derecesi?	x	0-25	
Sözlü Sınav			
Yapılan sunum başarılı mı?	x	0-5	
Soruları yanıtlama yetkinliği?	x	0-20	
Devam Durumu			
Öğrenci dönem içerisindeki raporlarını düzenli olarak hazırladı mı?	x	0-5	
Diğer Maddeler			
Toplam			

DANIŞMAN :
DANIŞMAN İMZASI: