Veri Madenciliği Uygulamalarında Özellik Seçimi İçin Finansal Değerlere Binning ve Five Number Summary Metotları ile Normalizasyon İşleminin Uygulanması

Ali Tunç[1], İlker Ülger[1]

[1] Kuveyttürk Katılım Bankası Konya AR-GE Merkezi, Konya

ali.tunc@kuveytturk.com.tr, ilker.ulger@kuveytturk.com.tr

Özet: Normalizasyon işlemi veri madenciliği konusunda önemli bir yer teşkil etmektedir. Makine öğrenmesi için gerekli olan farklı sınıflandırma ölçütlerinin birbirlerine karşı başarımlarının belirlenebilmesi için, gerçekleştirilecek uygulamalardan önce veri setinde; performans üzerinde doğrudan etkisi olan özelliklerin belirlenmesi, nihai sonuç üzerinde etkisi olmayan ya da minimum etkiye sahip özelliklerin ortaya çıkarılması için özellik seçimi "Feaute Selection" teknikleri kullanılır. Bu teknikler kullanılarak veri setindeki gerekli özelliklerin kullanılması ve doğru sonuçlara ulaşılması amaçlanmaktadır. Değişkenlerin ortalama ve varyansları birbirlerinden önemli ölçüde farklı olduğu taktirde büyük ortalama ve varyansa sahip değişkenlerin diğerleri üzerindeki baskısı daha fazla olur ve onların rollerini önemli ölçüde azaltır. Bu yüzden veri üzerinde normalizasyon işlemi yapılmalıdır. Binning ve Five Number Summary yöntemleri ile normalizasyon işlemleri yapılmıştır. Ve bu metotlardan elde edilen sonuçlar karşılaştırılarak hangi metodun daha uygun olduğu gözlemlenmeye çalışılmıştır.

Anahtar Sözcükler: Veri Madenciliği, Özellik Seçimi, Veri Normalizasyonu, Binning Metot, Five Number Summary Metot.

Implementation Of Normalization Process Using Binning and Five Number Summary Methods to Financial Value for Feature Selection in Data Mining Applications

Abstract: Normalization process has ultimate importance in data mining area. To determine releated performance between various classification measurements which required for machine learning, before application that will be performed in dataset; "Features Selection" technique is used to determination of direct effecting features on performance and to reveal features which have no or minimum effect on final result. The objective that using that techniques are using needed features and reaching accurate results. In the case of big differences between avarage and variation of variables, the variables which have bigger avarage and variation suppress on other variables and decrease others role importantly. Because of that normalization need to be take place on data. Binning and Five Number Summary method with normalization operations are performed. Which methods and comparing the results obtained from these methods have been tried to be observed to be more appropriate.

Keywords: Data Mining, Feature Selection, Data Normalization, Binning Metohods For Data Smoothing, The Five Number Summary.

1. Giriş

Bilgisayar ve bilgisayar teknolojileri hayatımızda çok önemli bir yer tutmaktadır. Her türlü bilgi ve veriler bilgisayarlarda tutulmakta ve bilgisayarlarda tutulan verilerin miktarı her geçen gün artış göstermektedir. Artan bu verilerin daha yararlı ve kullanılabilir hale getirilmesi için çeşitli yöntemler geliştirilmiştir.

Geliştirilen bu yöntemler verilerin birbiri ile olan ilişkileri üzerinden çeşitli işlemlerle sonuç üretmeye dayanır. Yaptığımız çalışmada da verilerin sınıflandırılabilmesi için verilerin birbirlerine göre ilişkilerinin incelenmesi gerekmektedir. Burada verilerin içerdiği bilgilere göre sonucu en çok ya da en az etkileyen özelliklerin bulunması çalışması yapılmıştır. Yapılan çalışma da belirli bir özellik üzerine normalizasyon yapılmış ve yapılmamış değerlere göre özelliğin sonuca etkisi gözlenmeye çalışılmıştır.

Bildirimizde veri madenciliği, özellik seçimi, binning metot ve the five number summary metot ile normalizasyon işlemi ve konularda yapmış olduğumuz çalışmalara değineceğiz.

2. Veri Madenciliği

Makine öğrenmesi alanında en önemli konulardan birisi verinin sınıflandırılmasıdır. Veri sınıflandırma işlemleri de veri madenciliği alanında değerlendirilir. Büyük miktardaki veriler içerisinden önemli olanlarını bulup çıkarmaya Veri Madenciliği denir. Veri madenciliği var olan bilgilerden anlamlı veri çıkarmayı hedefler. Veri madenciliği uygulamalarında alt yapı gereksinimi veri ambarı sayesinde sağlanır. Verilerin boyutlarından dolayı klasik veritabanı yöntemiyle işlenmesinin olanaksız olduğu durumlar için geliştirilmiştir. 1991 yılında ilk kez William H. Inmon tarafından ortaya atılan veri ambarı, yönetimin kararlarını desteklemek amacı ile çeşitli kaynaklardan elde ettikleri bilgileri zaman değişkeni kullanarak veri toplama olarak tanımlanmaktadır. Kısaca birçok veritabanından alınarak birleştirilen verilerin toplandığ depolardır. Veri ambarlarının özelliği kullanıcılara farklı detay düzeyleri sağlayabilmesidir. Detayın en alt düzeyi arşivlenen kayıtların kendisi ile ilgili iken, daha üst düzeyler zaman gibi daha fazla bilginin toplanması ile ilgilidir [1].

Veri Madenciliği, veri ambarlarında tutulan çok çeşitli verilere dayanarak daha önce keşfedilmemiş verileri ortaya çıkarmak, bunları karar vermek ve gerçekleştirmek için kullanma sürecidir. Bu tanımdan yararlanarak veri madenciliğinin aynı zamanda bir istatistiksel süreç olduğunu da söylemek mümkündür [2].

Son 20 yıldır Amerika Birleşik Devletleri'nde çeşitli veri madenciliği algoritmalarının gizli dinlemeden, vergi kaçakçılıklarının ortaya çıkartılmasına kadar çeşitli uygulamalarda kullanıldığı bilinmektedir. Kaynaklar incelendiğinde veri madenciliğinin en çok kullanıldığı alan olarak tıp, biyoloji ve genetik görülmektedir [3].

Faklı yerlerde ve farklı zamanlarda kliniklerde toplanan invaziv test verileri arasında yapılan veri madenciliği çalışmaları teşhiste %100 oranında doğruluk sağlamıştır [4].

Veri Madenciliği Sürecini 6 aşama olarak değerlendirebiliriz [5].

Veri Madenciliği Süreci

1. Veri temizleme
2. Veri bütünleştirme
3. Veri indirgeme
4. Veri dönüştürme
5. Veri madenciliği algoritmasını uygulama
6. Sonuçları sunum ve değerlendirme

Veri temizleme: Veri tabanında yer alan tutarsız ve hatalı verilere gürültü denir. Verilerdeki gürültüyü temizlemek için; eksik değer içeren kayıtlar atılabilir, kayıp değerlerin yerine sabit bir değer atanabilir, diğer verilerin ortalaması hesaplanarak kayıp veriler yerine bu değer yazılabilir, verilere uygun bir tahmin (karar ağacı, regresyon) yapılarak eksik veri yerine kullanılabilir [6].

Veri bütünleştirme: Farklı veri tabanlarından ya da veri kaynaklarından elde edilen verilerin birlikte değerlendirmeye alınabilmesi için farklı türdeki verilerin tek türe dönüştürülmesi işlemidir. Bunun en yaygın örneği cinsiyette görülmektedir. Çok fazla tipte tutulabilen bir veri olup, bir veri tabanında 0/1 olarak tutulurken diğer veri tabanında E/K veya Erkek/Kadın şeklinde tutulabilir. Bilginin keşfinde başarı verinin uyumuna da bağlı olmaktadır [6].

Veri indirgeme: Veri madenciliği uygulamalarında çözümlemeden elde edilecek sonucun değişmeyeceğine inanılıyorsa veri sayısı ya da değişkenlerin sayısı azaltılabilir. Veri indirgeme yöntemleri; veri sıkıştırma, örnekleme, genelleme, birleştirme veya veri küpü, boyut indirgeme [6].

Veri Dönüştürme: Verinin kullanılacak modele göre içeriğini koruyarak şeklinin dönüştürülmesi işlemidir. Dönüştürme işlemi kullanılacak modele uygun biçimde yapılmalıdır. Çünkü verinin gösterilmesinde kullanılacak model ve algoritma önemli bir rol oynamaktadır. Değişkenlerin ortalama ve varyansları birbirlerinden önemli ölçüde farklı olduğu taktirde büyük ortalama ve varyansa sahip değişkenlerin diğerleri üzerindeki baskısı daha fazla olur ve onların rollerini önemli ölçüde azaltır. Bu yüzden veri üzerinde normalizasyon işlemi yapılmalıdır [6].

Veri madenciliği algoritmasını uygulama: Veri hazır hale getirildikten sonra konuyla ilgili veri madenciliği algoritmaları uygulanır.

Sonuçları sunum ve değerlendirme: Algoritmalar uygulandıktan sonra, sonuçlar düzenlenerek ilgili yerlere sunulur. Örneğin hiyerarşik kümeleme yöntemi uygulanmış ise sonuçlar dendrogram grafiği sunulur.

2.1 Özellik Seçimi

Özellik seçimi (feature selection); özelliklerin alt kümelerini, doğruluktan ödün vermeden seçmektedir. İlgisiz verileri, gereksiz verileri silerek yüksek boyutu indirgemeyi hedeflemektedir. Amacı, gereksiz özellikleri çıkararak accuracy bulmayı hedefler.

Özellik seçimi herhangi bir veri madenciliği ürünü için bir gerekliliktir. Bir dataset içerisinde gerekli gereksiz birçok özellik barınabilmektedir. Var olan özellikler içerisinde sonucu en çok etkileyen yani sonuçla ilişkili olan özelliklerin belirlenmesine ihtiyaç vardır. Her hangi bir model oluştururken ilişkili verilerden hareket edilmesi gerekmektedir.

Genel olarak, özellik seçimi her öznitelik için bir puan hesaplama ve en iyi skorlar olan öznitelikleri seçerek çalışır. Üst skorları için eşik ayarlayabilirsiniz. Özellik seçimi, model, modelinde kullanılma olasılığı en yüksek olan bir dataset nesnesindeki öznitelikleri otomatik olarak seçmek için eğitilmiş veri modellerinden yararlanılır [11].

2.2 Normalizasyon

Verilerin veri bütünlüğünü bozacak şekilde, farklı ölçek ya da kod ile kaydedildiği durumlarda başvurulan bir yöntemdir. Buna örnek olarak maaş verisi, gelir, fiyat, tutar gibi finanssal verilerin ayrık değerler olarak sistemde tutulmasını örnek gösterebiliriz.

Veri dönüştürmede 3 yaklaşım kullanılabilir.

Ondalık Ölçekleme

Ondalık ölçekleme ile normalleştirmede, ele alınan değişkenin değerlerinin ondalık kısmı hareket ettirilerek normalleştirme gerçekleştirilir. Söz konusu ölçekleme, sayısal değerlerin -1 ile +1 arasında yer almalarını sağlayacak biçimde dönüştürülmesine karşılık gelir.

Hareket edecek ondalık nokta sayısı, değişkenin maksimum mutlak değerine bağlıdır. Ondalık ölçeklemenin formülü aşağıdaki şekildedir: Örneğin 900 maksimum değer ise, n=3 olacağından 900 sayısı 0,9 olarak normalleştirilir [7]. Bu işlem eşitlik 1'deki denklem yardımıyla hesaplanır.

$$v'(i) = V(i)/10^k \quad (1)$$

En küçük k değeri yani max(|v(i)|) < 1 için ölçeklenmiş değerdir.

Öncelikle maksimum |(v(i))| değeri veri seti içinden bulunur. Daha sonra ondalık nokta yeni, ölçeklendirilmiş ve mutlak değeri 1'den küçük oluncaya kadar hareket ettirilir. Sonrasında bölen diğer v(i)'lere uygulanır. Örneğin, en büyük değer 435 ve en küçük değer -834 iken özelliğin maksimum mutlak değeri 0.834 ve tüm v(i) için bölen 1 000 alınır. (k=3) [8].

Min-Max Normalleştirme

Min-Max normalleştirme ile orijinal veri üzerinde doğrusal bir dönüşüm yapılır. Bu yöntem aracılığıyla veriler genellikle [0-1] aralığına dönüştürülür. Min.-Max normalizasyon işlemi alan değerinin minimum değerden ne kadar büyük olduğuna bakar ve bu farkları sıralar [9]. Bu işlem eşitlik 2'deki denklem yardımıyla hesaplanır.

$$X^* = \frac{(X - \min(X))}{aralık(x)} = \frac{(X - \min(X))}{(maks(X) - \min(X))} \quad (2)$$

Min-max normalization: [$\min_A$, $\max_A$] to [new_min_A, new_max_A]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Ex. Let income [\$12,000, \$98,000] normalized to [0.0, 1.0]. Then \$73,000 is mapped to $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$

Şekil 1. Min-max normalizsyonu örnek denklem

Z-Score Standartlaştırma

Verileri dönüştürmek amacıyla kullanılan bir diğer yöntem Z-Score standartlaştırma yöntemi olarak bilinir. İstatistiksel veri dönüştürme teknikleri arasında yer alan ve en yaygın biçimde kullanılan bu yöntem, ele alınan verinin ortalama ve standart sapma değerlerini kullanır. Farklı tekniklerle gerçekleştirilebilen normalizasyon işlemi, veri boyutunun küçültülmesi amacıyla kullanılabileceği gibi, verilerle gerçekleştirilecek işlemlerin uygun aralıklara normalize edilmiş değerlerle yapılarak işlemlerin daha hızlı gerçekleştirilip ve daha anlamlı ve kolay yorumlanabilir sonuçlar almak amacı ile de kullanılabilir [10].

$$X^* = \frac{X - ortalama(X)}{standart\ sapma(X)} \quad (3)$$

eşitlik 3'deki denklem yardımıyla olan müşteri istekleri ile teknik gereksinimler ve bunların önem dereceleri belirlenmiştir.

2.2.1 Binning Metod (Binning Methods for Data Smoothing)

Eksik veri tamamlama, hatalı verileri düzeltme, tutarsız verileri kaldırma işlemine veri temizleme denir. Veri temizleme verinin işlemesi ve doğru sonuçların elde edilmesinde oldukça önemlidir.

Çalışmamızda binning metot işlemini Min ve Max değerleri bularak aykırı değerlerin temizlenmesi için kullandık. Projemizde maaş bilgisi gibi sürekli verilerin [0-1] aralığına yayılması gerekmektedir. Yalnız çok yüksek tutarda ve eksi tutarda maaş bilgisi sistemde yer almaktadır. Bunların tespit edilerek düzeltilmesi ve min max aralıklarının belirlenmesi için bu metodu kullandık [12].

Veriyi güncellemek için Binning Metodunda 3 yöntem kullanılmaktadır.

1. Ortalaması ile düzleştirme
2. Ortancası ile düzleştirme
3. Sınırları ile düzleştirme

Örnek DataSet :
4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
 Eşit Frekanslara ayrılırsa
- Bin 1: 4, 8, 9, 15
- Bin 2: 21, 21, 24, 25
- Bin 3: 26, 28, 29, 34
 Ortalama Değerlere Göre Düzeltme
- Bin 1: 9, 9, 9, 9
- Bin 2: 23, 23, 23, 23
- Bin 3: 29,29,29,29
 Sınırlara Göre Düzeltme:
- Bin 1: 4, 4, 4, 15
- Bin 2: 21, 21, 25, 25
- Bin 3: 26, 26, 26, 34

2.2.2 The Five Number Summary Metot

Beş sayı özeti var olan veri kümesi üzerin de

1. Minimum
2. First Quartile (Q1)
3. Median
4. Third Quartile (Q3)
5. Maximum

Beş adet bilgi bulunarak bu bilgilere göre verinin çeşitli amaçlar için kullanılması hedeflenmektedir. Bu değerlerin bulunma şekli var olan veri seti küçükten büyüğe sıralanır. 4 parça haline bölünür.

Q1 = n/4 ya da (n+1)/4 sıradaki değerdir.

Q3 = n* 3/4 ya da (n+1)*3/4 sıradaki değerdir.

IQR = Q3-Q1

LF = Q1 – (1.5 * IQR)

UF = Q3 + (1.5 * IQR)

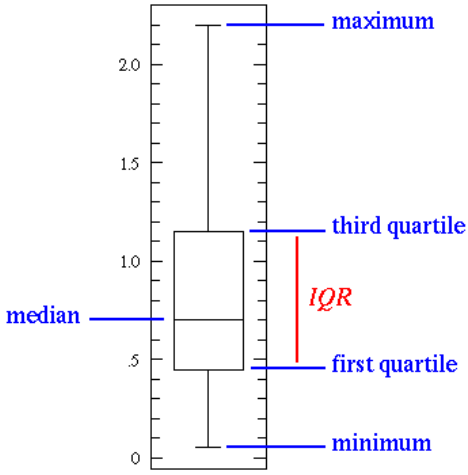
MIN = Listedeki >= LF deki ilk değer

MAX = Listedeki <= UF ilk değer

Median = Sıralamadaki ortada yer alan değer

Hesaplanan değerler sahip olunan minimum, maximum ve ortalama değer gibi bilgiler ile aykırı değerlerin belirlenmesini sağlar.

Bu beş değer belirlenerek var olan değerlerin min ve max bilgisine göre projemize ait normalizasyon çalışması yapılmıştır.



Şekil 2. Five Number Summary Box Plot Gösterimi

3. Yapılan Çalışma

Sistemde 10.000 e yakıt test datası üzerinden Salary, SectorId, ChildCount, SGKTypeId ve SelfEmployeeIncome bilgilerinin kazanım hesabını yaptık. Bu bilgileri 3 yönteme göre hesapladık. Bu yöntemleri karşılaştırarak sistem içerindeki özellik seçimi ile ilgili en uygun yöntemi bulmaya çalıştık.

3.1 Hiçbir normalizasyon işlemi yapmadan verilen bilgilerin hesaplanmasıdır. Bu işlem sütundaki değerlerin hiçbir işlem yapılmadan özellik seçimi fonksiyonlarına sokularak işlem yapılmasıdır. Değerler yüksek olduğu, varyans ve standart sapmaları yüksek sonuçlar ürettiği için baskın sonuçlar özellik sonuçları ortaya çıkmıştır. Ve buradaki değerlerin normal [0-1] aralığına dönüştürülmesine ihtiyaç olduğu gözlenmiştir.

3.2 Bining yöntemi aracılığı ile min max değerlerini bularak [0-1] arası normalizasyonun yapılması işlemidir. Bu işlem için ilgili kolon üzerindeki veriler ASC olarak sıralandı. Her 100 kayıt bir grup olacak şekilde gruplandırıldı. %20 değer içermeyen gruplar hesaplamaya dahil edilmedi. Her bir grubun ortalama değeri hesaplandı ve böylece grup ortalamalarına göre min – max değerler bulundu. Bulunan min ve max değerler ile normaliazasyon hesaplaması yapılmaya çalışıldı. Gerekli veri setini gruplara bölerek min ve max değerlerini belirleyen sorgu aşağıdaki gibidir.

```
SELECT
        @MinSalaryValue = MIN(Salary),
        @MaxSalaryValue = MAX(Salary)
FROM (SELECT
        SalaryGroup,
        AVG(Salary) AS Salary
FROM (SELECT
        ROW_NUMBER() OVER (ORDER
BY Salary ASC) AS 'Row',
        SALARY,
        ROW_NUMBER() OVER (ORDER
BY Salary ASC) / 100 AS SalaryGroup
FROM INF.Customer WITH (NOLOCK))
AS TBL
GROUP BY SalaryGroup
HAVING COUNT(SalaryGroup) > 20)
AS RST1
```

Burada elde edilen min max değerleri “Min Max Normalizasyon” yöntemi kullanılarak [0-1] aralığına getirme işlemine tabi tutuldu. Ve gerekli sonuçlar hesaplanarak kaydedilmiştir.

3.3 Five Number Summary yöntemin aracılığı ile min max değerlerinin bulunması ve aykırı değerlerin temizlenerek [0-1] arası normaliasyonun yapılması işlemidir. Bu işlem için ilgili kolon üzerindeki veriler ASC olarak sıralandı. Sıralanan kayıt sayısı 5 parçaya bölündü. Q1 değeri için n*1/5 sırasındaki kayıtın değeri alındı. Q3 için n*4/5 sıradaki kayıt alındı. IQR hesaplandı ve bu verilere göre olması gereken aralık bulundu. Bulunana aralığın içindeki Min ve Max değerler hesaplandı. Min max değerlerini hesaplayan sorgu aşağıdaki gibidir.

```
SELECT @MinSalaryValue = ( Q1 - (1.5 *
QRC)) ,  @MaxSalaryValue = ( Q3 + (1.5*
QRC))  FROM (
SELECT        MIN(Salary)  AS  Q1  ,
MAX(Salary) AS Q3,  (MAX(Salary) -
MIN(Salary)) AS QRC FROM (
SELECT ROW_NUMBER() OVER
(ORDER BY Salary ASC) AS 'Row1',
SALARY
FROM INF.Customer WITH (NOLOCK)
)AS TABLESALARY
WHERE
Row1 IN (@RecordCount *
@Q1/@QRC,@RecordCount* @Q3/@QRC)
)  AS TBLSAL

SET @MinSalaryValue =  ( SELECT TOP 1
Salary FROM INF.Customer WITH
(NOLOCK) WHERE Salary >=
@MinSalaryValue ORDER BY Salary
ASC);
SET @MaxSalaryValue =  ( SELECT TOP 1
Salary FROM INF.Customer WITH
(NOLOCK) WHERE Salary <=
@MaxSalaryValue ORDER BY Salary
DESC);
```

Burada elde edilen min max değerleri “Min Max Normalizasyon” yöntemi kullanılarak [0-1] aralığına getirme işlemine tabi tutulur. Ve gerekli sonuçlar hesaplanarak

kaydedilmiştir. [0-1] aralığı değiştirilmek istenmesi durumunda hazırlanmış sql ifade de sadece parametre değiştirmesi yeterli olacak şekilde sistem tasarlanmıştır. Bu sorguda @NewMinValue ve @NewMaxValue değerleri [0-1] yerine [X-Y] verilirse normalizasyon o aralığa göre yapılmaktadır.

((@NewMaxValue-@NewMinValue) * ((@ColumnValue - @MinValue) / (@MaxValue-@MinValue))) + @NewMinValue ;

Yapılan çalışmalarda elde edilen sonuçlar tablosal olarak da gösterilmiştir. Aşağıdaki tablolarda normalize edilmiş ve edilmemiş olan alanların diğer alanlara göre hesaplanmış olan Gain Ratio değerleri görünmektedir. Five Number Summary Yöntemi ve Bining yöntemi uygulanan hesaplamalar ayrı ayrı tablo olarak konulmuş ve bu değerlerin karşılaştırılması sağlanmıştır.

Five Number Summary metodunda bulunan min ve max değer aralığının binning metodunda bulunan değer aralığına göre daha küçük olması ortaya çıkan normalizasyon ve özellik değerlerinin sonucu tablolarda görünmektedir. Tablo 1'de Five Number Summary, Tablo 2'de de Binning Metot sonuçları gösterilmektedir.

SelfEmployeeIncome	0.009591
SectorId	0.008451
ChildCount	0.008161
Salary	0.007634
SelfEmployeeIncome Normalization	0.006731
SGKTypeId	0.006575
SalaryNormalization	0.006421

Tablo 1. Five Number Summary Yöntemi Normalleştirilimiş ve Normalleştirilmemiş kayıtların Hesaplanmış Gain Ratio Değerleri

SelfEmployeeIncome	0.009591
SelfEmployeeIncome Normalization	0.009286
SectorId	0.008451
ChildCount	0.008161
Salary	0.007634
SalaryNormalization	0.007618
SGKTypeId	0.006575

Tablo 2. Binning Yöntemi Normalleştirilimiş ve Normalleştirilmemiş kayıtların Hesaplanmış Gain Ratio Değerleri

4. Sonuç ve Değerlendirme

Yapılan çalışmalarda normalizasyon yapılmamış verilerin özellik hesaplarında yüksek değerler olarak çıktığı için karar verme aşamalarında bazı olumsuz değerlendirmelere yol açabileceği gözlenmiştir. SelfEmployeeIncome ve SectorId karşılaştırmasını yapacak olursak nomralizasyon yapılmadığı taktirde SectorId bilgisinin değeri SelfEmployeeIncome değerinden düşük çıkmaktadır. Ama normalizasyon yapıldığı taktide SectorId , ChlidCount gibi alanlar normalizasyon yapılmış SelfEmployeeIncome alanından daha yüksek değere sahip olduğu gözlenmiştir.

0 - 50.000 aralığında yer alan Salary bilgilerinin Bining metodu ile hesaplanmasında min değer 0 max değer 42.000 gibi bir sonuç bulunmuştur.

Aynı Salary bilgisi Five Number Summary metodu ile hesaplandığında ise min değer 0 max değer ise 7.500 bulunmaktadır.

Yapılan çalışmalar görülmektedir ki frekanssal dağılımın yoğun olduğu aralıkların MİN ve MAX değerlerinin bulunmasın da Five Number Summary Metodu Binning

Metoduna göre daha doğru sonuçlar verdiği tespit edilmiştir. Eşit grupsal dağılımın yapılabildiği aralıklarda ise Binning Metodun daha doğru sonuçlar verdiği gözlenmiştir.

5. Kaynaklar

[1] Murray J. Mackinnon ve Ned Glick, 'Data Mining and Knowledge Discovery in Databases- An Overview', **J.Statists., Vol.41, No.3 s.260**, (1999).

[2] K. Yaralıoğlu, Veri Madenciliği, **http://www.deu.edu.tr/userweb/k.yaraliogl u**, (Mayıs, 2013)**.**

[3] S. Savaş, N. Topaloğlu, M. Yılmaz, ' Ver. Madenciliği ve Türkiye' deki Uygulama Örnekleri, İstanbul Ticaret Üniversitesi, **Fen Bilimleri Dergisi, 21, 1-23,**(2012)**.**

[4] A.Kusiak, K.H. Kernstine, J.A.Kern, K.A.McLaughlin and T.L.Tseng: **Medical and Engineering Case Studies** (2000)**.**

[5] Özge Kaplan, Gizem Gözen "ORACLE DATA MINER" ile mantarların zehirliliği üzerine bir veri madenciliği uygulaması, İstanbul Teknik Üniversitesi Fen Edebiyat Fakültesi Mat. Mühendisliği programı, (2010).

[6] E. Coşku, Veri Madenciliği, **http://ab.org.tr/ab13/bildiri/175.pdf**, (Mayıs, 2013).

[7] Oğuzlar A. "Veri Ön İşleme", **Ege Üniversitesi İktisadi ve İdari Bilimler** , (2003) **Fakültesi Dergisi, Sayı 21, Temmuz-Aralık** 2003, s.73, **http://iibf.erciyes.edu.tr/dergi/sayi21/ aoguzlar.pdf** , (2012.)

[8] Kantardzic, M.M. ve Zurada, J. Next Generation of Data-Mining Applications. New Jersey: **Institute of Electrical and Electronics Engineers Inc.** (2005).

[9] Larose, D.T. Discovering Knowledge In Data: An Introduction to Data Mining. New Jersey: **John Wiley and Sons Inc.** (2005).

[10] Khemka, A., A Colloborative Predictive Data Mining Model, Yayınlanmamış Yüksek Lisans Tezi, **Faculty of University of Missouri-Kansas City, Missouri** (2003) .

[11] **https://technet.microsoft.com/tr-tr/library/ms175382(v=sql.105).aspx**

[12] Doç. Dr. Suat Özdemir - **Veri Madenciliği Ders Notları** (2010).