**Hafta-4**

İstatistik ve olasılık, sayısal analiz ve özellikle veri madenciliği dersi kapsamında elde ettiğiniz kazanımları burada değerlendirmenizi bekliyoruz. Veri seti içerisinde yer alan özelliklerin alt/üst aralıkları, kritik seviyelerin neler olabileceği, günün zaman dilimine göre, çalışma günlerine göre, tatil günlerine göre veya mevsimsel olarak nasıl bir değişime sahip olduğu ayrıca özelliklerin birbirleri arasındaki ilişkilerin olup olmadığının analizi beklenmektedir.

**Veri:** İşlenmemiş, yorum yapmaya imkân verecek düzeyde birbiriyle ilişkilendirilmiş ham kayıtlar.

**Veri Madenciliği Nedir?**

Basit bir tanım olarak veri madenciliği büyük ölçekli veriler arasında bilgiye ulaşma veya bilgiyi madenleme işidir. Büyük veri yığınları içerisinden gelecekle ilgili tahminde bulunabilmemizi sağlayabilecek bağıntıların bilgisayar programı kullanarak aranmasıdır. Bunun dışında bilgi keşfi, bilgi madenciliği, bilgi çıkarımı, veri/model analizi, veri arkeolojisi… Veri Madenciliği için kullanılan alternatif terimlerdir. Kısacası büyük veri yığınları arasında görülmeyen bilgiyi ortaya çıkarma işlemidir.

**Kullanılacak Algoritmalar**

Projemiz için çeşitli sonuçları elde edebilmek için madenciliğinde birçok algoritma bulunmaktadır bizim kullanacağımız algoritmalar Apriori ve C 4.5’dur. Bu algoritmaları kullanabilmek için yaygın olarak kullanılan weka programı kullanılacaktır.

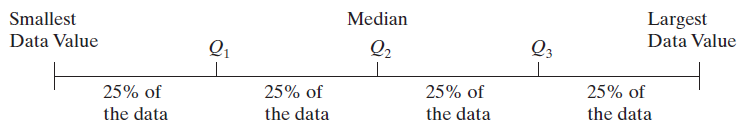
**Uç değer Hesabı**

Uç değer hesabı verilerin dağılımlarını anlayabilmek ve yorumlamalar yapabilmek adına çok önemlidir. Uç değerler genel anlamda, veri setinde bulunan diğer değerler tarafından ifade edilen eğilimden büyük ölçüde farklı olan veri noktalarıdır. Başka bir deyişle, uç değerler diğer değerlerin dışında kalan değerlerdir. Bunları tespit etmek için çeşitli yazılımlar ve yöntemler olduğu gibi, veri tablolarında ve grafiklere bakarak ta anlaşılabilirler.

**Uç Değer Hesabı**

[Uç Değerler](http://mustafaakca.com/tag/uc-degerler/) Nasıl Belirlenir? Ekstrem bir değeri belirlemenin en kolay yolu şudur.

Ortalamalar veriyi ikiye bölen değerdir, Kartiller ise 4 bölen değerlerdir.

[](http://mustafaakca.com/wp-content/uploads/2014/08/kartil.png)

2 numaralı [kartil](http://mustafaakca.com/tag/kartil/" \o "kartil) aynı zamanda ortalamadır. Çeyrekler arası mesafeyi kontrol etmek için şu formül kullanılır.

IQR=Q3-Q1

ve bir değerin [outlier](http://mustafaakca.com/tag/outlier/" \o "outlier) ([uç değer](http://mustafaakca.com/tag/uc-deger/)) olup olmadığına ise şu şekilde bakılır.

Alt [UçDeğer](http://mustafaakca.com/tag/uc-deger/" \o "Uç Değer)  = Q1 – 1.5(IQR)  
Üst [Uç Değer](http://mustafaakca.com/tag/uc-deger/) = Q3 + 1.5(IQR)

Eğer bu değerleri aşan bir değer varsa, [outlier](http://mustafaakca.com/tag/outlier/" \o "outlier)([uç değer](http://mustafaakca.com/tag/uc-deger/))’dir diyebiliriz.

Bu değerler Boxplot diyagramlarını hazırlamakta bize gerekli olacak değerlerdir. Q1, Q2([Ortalama](http://mustafaakca.com/tag/ortalama/)) ve Q3’ü bilinen bir dağılımın boxplot diyagramı alttaki gibi çizilebilir.