T.C.
DÜZCE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

OTOMOTİV SEKTÖRÜNDE KALİTE KONTOL SÜRECİNDE VERİ MADENCİLİĞİ YÖNTEMLERİ İLE KARAR DESTEK SİSTEMİ UYGULAMASI

HİKMET CANLI

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

DANIŞMAN
YRD. DOÇ. DR. Sinan TOKLU

DÜZCE, 2017

T.C.
DÜZCE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

OTOMOTİV SEKTÖRÜNDE KALİTE KOTROL SÜRECİNDE VERİ MADENCİLİĞİ YÖNTEMLERİ İLE KARAR DESTEK SİSTEMİ UYGULAMASI

Hikmet CANLI tarafından hazırlanan tez çalışması aşağıdaki jüri tarafından Düzce Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Tez Danışmanı

Yrd. Doç. Dr. Sinan TOKLU

Düzce Üniversitesi

Jüri Üyeleri

Yrd. Doç. Dr. Mehmet ŞİMŞEK

Düzce Üniversitesi

Jüri Üyeleri

Yrd. Doç. Dr. İbrahim Alper DOĞRU

Gazi Üniversitesi

Tez Savunma Tarihi: 29/12/2017

BEYAN

Bu tez çalışmasının kendi çalışmam olduğunu, tezin planlanmasından yazımına kadar bütün aşamalarda etik dışı davranışımın olmadığını, bu tezdeki bütün bilgileri akademik ve etik kurallar içinde elde ettiğimi, bu tez çalışmasıyla elde edilmeyen bütün bilgi ve yorumlara kaynak gösterdiğimi ve bu kaynakları da kaynaklar listesine aldığımı, yine bu tezin çalışılması ve yazımı sırasında patent ve telif haklarını ihlal edici bir davranışımın olmadığını beyan ederim.

29 Aralık 2017

Hikmet CANLI

TEŞEKKÜR

Yüksek lisans öğrenimimde ve bu tezin hazırlanmasında gösterdiği her türlü destek ve yardımdan dolayı çok değerli hocam Yrd.Doç.Dr. Sinan TOKLU'ya en içten dileklerimle teşekkür ederim.

Tez çalışmam boyunca değerli katkılarından dolayı Özgür DEMİR'e de şükranlarımı sunarım.

Bu çalışma boyunca yardımlarını ve desteklerini esirgemeyen sevgili aileme ve çalışma arkadaşlarıma sonsuz teşekkürlerimi sunarım.

29 Aralık 2017 **Hikmet CANLI**

İÇİNDEKİLER

Sayfa No

ŞEKİL LİSTESİ

Sayfa No

ÇİZELGE LİSTESİ

Sayfa No

KISALTMALAR

CRISP	Çapraz-endüstri standart işlem
RF	Random Forest
SMO	Sequential minimal optimization

ÖZET

OTOMOTİV SEKTÖRÜNDE KALİTE KONTROL SÜRECİNDE VERİ MADENCİLİĞİ YÖNTEMLERİ İLE KARAR DESTEK SİSTEMİ UYGULAMASI

Hikmet CANLI
Düzce Üniversitesi
Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı
Yüksek Lisans Tezi
Danışman: Yrd. Doç. Dr. Sinan TOKLU
Aralık 2017, 48 Sayfa

Günümüzde otomotiv sektörü, gelişmiş ve hatta gelişmekte olan ülkeler için "anahtar" sektör rolündedir. Güçlü bir otomotiv sektörü, sanayileşmiş ülkelerin ortak özelliklerinden biri olarak gözümüze çarpmaktadır. Bu sektörde üretim birçok süreçten oluşmaktadır. Bu süreçlerin en önemli olanlarından biri de kalite kontroldür. Bu alanda ölçüm verileri çok fazladır ve verilerin hacmi arttıkça insanların anladığı oran azalmaktadır. Varyasyonlar kalitenin düşmanıdır ve her şeyde varyasyon bulunmaktadır. Bu tez çalışmasında veri madenciliği yöntemlerinden olan sınıflandırma algoritmaları ile kalite kontrol sürecinde bir karar destek sistemi uygulaması yapılmıştır. Bu çalışma hazırlanırken veri madenciliği için Çapraz Endüstri Standart İşlem Modeli (CRISP) kullanılmıştır. Çalışmada sınıflandırma algoritmaları sonuçların performansları Çapraz Geçerleme ve Hold-Out yöntemleri ile karşılaştırılmıştır. Çapraz geçerleme katı olarak 5 kat ve 10 kat çapraz geçerleme katı kullanılmıştır. Hold-Out yöntemi ile de %40-%60, %25-75, %20-%80 ayrım oranlarına sahip sırasıyla test ve eğitim veri setine ayrılmıştır. Karşılaştırma sonucunda karar ağacı ile kurulan modeller diğer modellerden daha iyi sonuç vermiştir. En iyi performansı gösteren C4.5 karar ağacı algoritmasının doğruluk oranı yaklaşık 0.87'dir. Yine başka bir karar ağacı olan Random Forest algoritması da yüksek bir doğruluk oranına ulaşsa da zaman performansı olarak geride kalmıştır. Bu iki algoritmayı performans olarak NaiveBayes ve SMO algoritmaları izlemektedir. Bu çalışmada ek olarak veri madenciliği yöntemlerinden biri olan veri görselleştirme teknikleri kullanılarak kalite analizi için bir uygulamaya da yer verilmiştir.

Anahtar sözcükler: Kalite kontrol, Karar destek, Üretim, Veri görselleştirme, Veri madenciliği.

ABSTRACT

APPLICATION OF DECISION SUPPORT SYSTEM WITH DATA MINING METHODS IN AUTOMOTIVE SECTOR IN QUALITY CONTROL

Hikmet CANLI
Düzce University
Graduate School of Natural and Applied Sciences, Department of Computer Engineering
Master's Thesis
Supervisor: Assist. Prof. Dr. Sinan TOKLU
December 2017, 48 pages

Today, the automotive sector is the "key" sector for developed and even developing countries. A strong automotive sector is striking as one of the common features of industrialized countries. Production in this sector consists of many processes. One of the most important of these processes is quality control. The measurement data in this area is very large and as the volume of data increases, the rate that people understand is reduced. Variations are the enemy of quality and there is variation in everything. In this thesis study, a decision support system is applied in the quality control process with classification algorithms which are data mining methods. While this work was underway, the Cross Industry Standardized Processing Model (CRISP) was used for data mining. The performance of the results of the classification algorithms in the study was compared with the Cross Validation and Hold-Out methods. With the Hold-Out method, the test and training data set is divided into 40% -60%, 25-75%, 20% -80% discrimination ratios respectively. As a result of the comparison, the models established with the decision tree gave better results than the other models. The best performing C4.5 decision tree algorithm has an accuracy rate of about 0.87. Yet another decision tree, the Random Forest algorithm, has reached a high accuracy rate, but is still out of time performance. These two algorithms are followed by Naive Bayes and SMO algorithms in performance. In this study, an application for quality analysis using data visualization techniques, which is one of the data mining methods, is also included.

Keywords: Data mining, Data visualization, Decision support, Production, Quality control.

1. GİRİŞ

Kalite kontrolün amacı, tüketici isteklerinin ve işletmenin genel gayesini birlikte muhtemel en ekonomik seviyede karşılayabilecek ürünün üretilmesini sağlayacak plan ve programların geliştirilerek uygulanması ve etkin bir şekilde yürütülmesini sağlamaktır. Bir ürün üretim aşamasında pek çok operasyondan geçmektedir. Bu operasyonların her biri belli kontrol planlarına sahiptir. Kontrol planları üretilecek parçanın ilgili operasyonuna ait ölçüm nominal ölçüm değerlerini içermektedir. Bir kontrol planında tek bir operasyon için yüzlerce ölçüm değeri gerekebilir. Seri üretim yapan otomotiv sektöründeki işletmelerde ürün sayısının çok olmasından dolayı ölçüm değerleri çok büyük veri setleri oluşturmaktadır. Veri seti büyüdükçe bunları anlamak analiz etmek zorlaşır ve zaman kaybettirmektedir. Ayrıca temel istatistik ve akıl yürütme yöntemleriyle yapılan analizler bize üretim gerçekleştikten sonra sonuç vermektedir ve belli bir kural ve tahmin oluşturamamaktadır. Bu tez çalışmasının amacı uygun veri madenciliği yöntemleri kullanarak kalite sürecinin daha iyi daha hızlı bir şekilde anlaşılmasını sağlamaktır. Ayrıca ürünlerin tamamını incelemek yerine belirli zaman aralıklarında prosesi yeterince temsil edebilecek nitelikte örneklemler çekilir. Amaç bu veri içinde saklı, gelecekle ilgili tahmin yapmakta kullanılabilecek kural ve bağıntıların çıkarılmasıdır.

Kalite kontrol veri madenciliğinin uygulama alanlarından bir tanesidir. Veri tabanı üzerinden elde edilen veriler üzerinde uygulanan kalite kontrol yöntemleriyle, kalite düzeyinin istenilen standartlara uygun olup olmadığı araştırılır. Eğer kalite düzeyi istenilen standartlara uygun değilse, kaliteyi istenilen seviyeye çıkartmak amacıyla çeşitli önlemler alınır.

Kalite kontrolde veri madenciliğinden yararlanılması, veriye daha çabuk ve kolay ulaşılmasını, dolayısıyla zaman ve maliyetten tasarruf edilmesini sağlar. Literatür taraması yaparak kalite kontrol sürecinde yapılan başlıca veri madenciliği çalışmalarına bakacak olursak;

Deng ve Wang tarafından zaman serisi veri madenciliği metodolojisine dayanarak, su kalitesinde zaman serisi verileri için yeni ve genel bir analiz çerçevesi önerisinde bulunulan bir çalışma yapılmıştır. Bu çalışma iki bölümden oluşmaktadır; uygulama bileşenleri ve su kalitesi verilerindeki zaman serisi veri madenciliğinin ortak görevleri. İlk bölümde, zaman serilerini iki boyutlu normal bulutlara parçalamayı ve granüle seviyedeki benzerlikleri hesaplamayı önermişlerdir. İkinci kısımda benzerlik matrisi ile su kalitesi zaman serisi örnek verileri ile benzerlik araştırması, anormallik tespiti ve model bulma çalışmaları yapılmıştır. Çin'in Yangtze Nehri'nin üst menzilindeki beş izleme istasyonundan toplanan haftalık Dissolve Oksijen zaman serisi verilerine ilişkin bir vaka çalışmasını incelemişlerdir. Deneysel sonuçlar, önerilen analiz çerçevesinin, su kalitesindeki tarihsel zaman serisi verilerinden gizli ve değerli bilgiyi keşfetmek için uygulanabilir ve etkili bir yöntem olduğunu göstermiştir [1].

Baykasoğlu yaptığı çalışmada veri madenciliği ve uygulama alanlarını bahsetmiş ve daha sonra da çimento sektöründe yaptığı bir uygulamayı anlatmıştır. Basma dayanıklılığı en önemli çimento özelliğidir, öyle ki kalite kontrol için ana parametredir. Basma dayanıklılığının belirlenmesi için standart “28 gün basma dayanıklılığı testi” yaygın olarak kullanılır. Bu test çimento üretimi sürecinde her partiden alınan numunelerin 28 gün bekletilerek basma mukavemetin deneysel olarak belirlenmesini içerir. Fakat çimento basma dayanıklılığının deneysel sonuçlarının elde edilmesi için 28 gün beklemek endüstri için uzun bir zamandır. Bu nedenle, basma mukavemetinin daha hızlı belirlenmesi çimento endüstrisi için bir ihtiyaçtır ve araştırmacıların ilgisini hak etmektedir. Çalışmada Portland kompozit çimentosunun basma mukavemetini tahmin etmek için gen denklem programlama ve yapay sinir ağları olarak bilinen iki yeni nesil veri madenciliği yöntemi ve regresyon analizi olarak bilinen klasik bir veri madenciliği yöntemi kullanılarak bu yöntemlerin performansları karşılaştırılmıştır. Yapılan çalışma sonucunda yapay zekâ temelli yöntemlerin daha iyi sonuç verdiği gözlenmiştir. Özellikle gen denklem programlama diğer yöntemlerden daha iyi sonuç vermiştir [2].

Glawar ve arkadaşları yaptığı bir çalışmada veri madenciliği ile desteklenen kalite odaklı bakım planlaması üzerinde uygulamasını anlatmıştır. Çalışmaya göre doğru zamanda gerçekleştirilen uygun bakım tedbirleri, modern imalat sistemlerinde tesisin kullanılabilirliğini, ürün kalitesini ve süreç verimliliğini güvence altına almak için önemli

bir etkendir. Kurulan bakım stratejileri, çoğu kez, bu güçlü ilişkili yönleri birleştirmede yetersiz kalmıştır. Bütüncül bir şekilde tahmin edebilecek durumda değildirler ve bu nedenle gereksiz yüksek bakım çalışmaları, zaman kaybına, kalite ve erişilebilirlik bozukluklarının ortaya çıkmasına neden olmuştur. Bakım planlaması için bütüncül ve öngörücü bir yaklaşım gerçekleştirmek için, çeşitli verilerin "sebep sonuç" tutarlılıkları ile tutarlı bir şekilde derlenmesi ve ilişkilendirilmesi için bir yöntem önerilmektedir. Bileşen seviyesindeki üretim tesislerini parçalayarak, veri madenciliği yöntemlerini kullanarak durum izleme verilerini, veri yıpranmasını, kaliteyi ve üretim verilerini birbirine bağlamak için bir temel oluşturulur. Bu çerçeve, kritik bakım koşulların belirlenmesini, hata momentlerinin ve kalite sapmalarının öngörülmesini sağlamaktadır [3].

Harding ve arkadaşları imalatta veri madenciliğini konusunda detaylı bir araştırmada bulunmuşlardır. Bu araştırma, veri madenciliği üretim mühendisliği uygulamaları, özellikle üretim süreçleri, operasyonlar, arıza tespiti, bakım, karar destek ve ürün kalitesi iyileştirme konularını incelemiştir. Araştırmalarında, genel olarak veri madenciliği alanını tartışmak yerine, veri madenciliğinin imalat sanayii ile alakalı olduğunu göstermeye çalışmışlardır. Bu araştırmada veri madenciliği üretiminde sayısız uygulama incelenmiştir. Son yıllarda arıza tespiti, kalite iyileştirme, üretim sistemleri ve mühendislik tasarımı gibi bazı imalat alanlarındaki yayın sayısının önemli bir artışı vardır. Diğer alanlar nispeten daha az önem görmektedir. Ayrıca araştırmada veri madenciliğinin imalat sanayinde büyümesini sağlamak için veri temizleme için daha genel bir sürecin gerekli olduğuna kanısına varılmıştır [4].

Kamal tarafından yapılan bir araştırmada üretimdeki kalite kontrol sürecini geliştirmek için uygulanan veri madenciliği yaklaşımları anlatılmıştır. Verilerin hacmi arttıkça kaçınılmaz olarak insanın anladığı oran küçülmektedir. Çeşitlilik ve ihtimal kalitenin düşmanıdır. Ürün kalitesi, herhangi bir işlem için odak noktası olmalıdır. Uygun veri madenciliği araçlarını ve istatistiksel akıl yürütme kavramını kullanarak, yöneticiler ve çalışanlar süreçlerinin daha iyi anlaşılmasını sağlamışlardır. Kamal veri madenciliğinin yanında SPC'sinin de kalitedeki varyasyonların anlaşılmasında önemli bir rol aldığını çalışmasında anlatmıştır. Sonuç olarak kalite kontrol sürecinde veri madenciliği konsepti ve teknikleri, SPC tasarımı ve performansı verilerdeki kalıpları aramak ve iyileştirmek için genel bir bakış oluşturmaktadır [5].

Khan ve arkadaşları tarafından yapılan bir çalışmada üretimde verimli kalite kontrolü için istatistiki veri madenciliğini yöntemleri anlatılmıştır. Makinelerin yaygın kullanımı, esnek / yeniden yapılandırılabilir üretim ve tamamen otomatikleştirilmiş fabrikalara geçiş, üretim sürecinde kaydedilen bilgilerin akıllıca kullanılmasını zorunlu kılmaktadır. Modern üretim süreçleri, sürecin farklı aşamalarında, örneğin sensör ölçümleri, makine okumaları vb. Boyunca Terabaytlık bilgi üretir ve bu büyük veri setinin ana katkısı, farklı kalite kontrol süreçleridir. Çalışmada imalat verilerinden değerli bilgiler elde etmek için bir yöntem öngörülmüştür. Önerilen yönteme göre, İstatistikte Genetik Algoritma için bir performans fonksiyonu olarak olasılıkların ve olasılık ilkelerinin uzatılmasının karşılaştırılmasına dayanmaktadır. Yapılan çalışma sonucunda 0,0095 civarında hata ihtimaline izin vererek QC7 (UT) kaynaklarının yaklaşık% 98'ini iyileştirilebildiği sonucuna varılmıştır. Ancak, endüstriyel kalite standartlarına göre bu rakam 0,00025'i geçmemelidir. Bu kalite standartlarını yakalamak için de farklı yöntemler düşünmüşlerdir [6].

Chen ve arkadaşları tarafından yapılan bir çalışmada üretim endüstrinde kalite kontrol tasarımı için veri madenciliği kullanımını anlatmışlardır. Çalışmada yarı iletken tesislerin üretim sürecindeki temel tutarsızlık nedenlerini keşfetmekte kullanılan iki veri madenciliği sınıflandırma analizinin(Karar Ağaçları ve Bayes Algoritması) doğruluğu karşılaştırılmıştır. Çalışmada dört özellik incelenmiştir; İnsan, Makine, Malzeme ve Yönetim. Amaç en kısa sürede en iyi ektiyi verecek makinayı tespit etmektir. Elde edilen sonuçlara göre Karar ağacı algoritması, yarı iletken ambalaj endüstrisinde kalite problemlerini analiz etmek için Bayes algoritmasından daha etkilidir ve uygundur sonucuna varılmıştır [7].

Ferreiro tarafından yapılan çalışmada havacılık endüstrisinde delme işleminde çapak algılamak için yapılan kalite kontrol sürecinde veri madenciliği incelenmiştir. Basit bir deney tasarımı ve veri madenciliği tekniklerinden, özellikle değişkenlerin seçimi ve makine öğrenme algoritmalarından, delme işlemi sırasında çapak saptanması için bir model geliştirilmiş. Model makine iç sinyaline ve sürecin koşullarının belirli parametrelerine dayanır bu nedenle uygulaması daha kolay olmuş ve harici sensörler kullanılmamıştır. Delme işlemi sırasında çapak oluştuğunda on-line tespit etmek için bir izleme sistemi oluşturulmuştur. Ve ikinci olarak delme işleminin hangi parametrelerinde

çapak oluşup oluşmadığı tanımlanır. Sonuçlarla ilgili olarak, hemen hemen tüm gelişmiş modeller mevcut matematiksel modelden daha yüksek bir doğruluk sağlar. Dahası, Naive Bayes'e dayanan nihai model için doğruluk% 95 ve standart sapma 0'a eşittir, yani çok kararlı bir model olduğu anlamına gelir. Bu noktada, modelin kötü bir tahminde bulunduğu durumların çoğunda, çapağın havai sınırlara (127 lm) çok yakın olması ve bunun hangi rüzgâr oranının belirlenmesini daha zor hale getirdiği de belirtilmiştir. Dikkate alınması gereken diğer bir husus, delinmiş delik çapak olsun veya olmasın, tahmini hatanın aynı önemi bulunmamasıdır. Bu vakaların tespiti için model geliştirilerek, çapakların incelenme sayısı önemli ölçüde azaltılmış [8].

İncelenen bu çalışmalarda genel amaç kalite kontrol sürecinin zaman olarak azaltılması, kaliteyi bozan durumların tespiti ve karar verme yetisinin kolaylaştırılmasıdır. Bu araştırmalar açıkça göstermektedir ki veri madenciliğinin kalite kontrol sürecinde çok önemli bir yeri bulunmaktadır. Ayrıca genel kanı olan tahmin için sınıflandırma algoritmaları kullanıldığı bu çalışma ile doğrulanmıştır. Bu çalışmalar neticesinde yapılacak çalışmada veri madenciliği tekniklerinden sınıflandırma algoritmaları kullanılmasına karar verilmiştir ve bu çalışmada C4.5, Random Forest, SMO ve Bayes algoritmaları üzerine çalışma yapılmıştır.

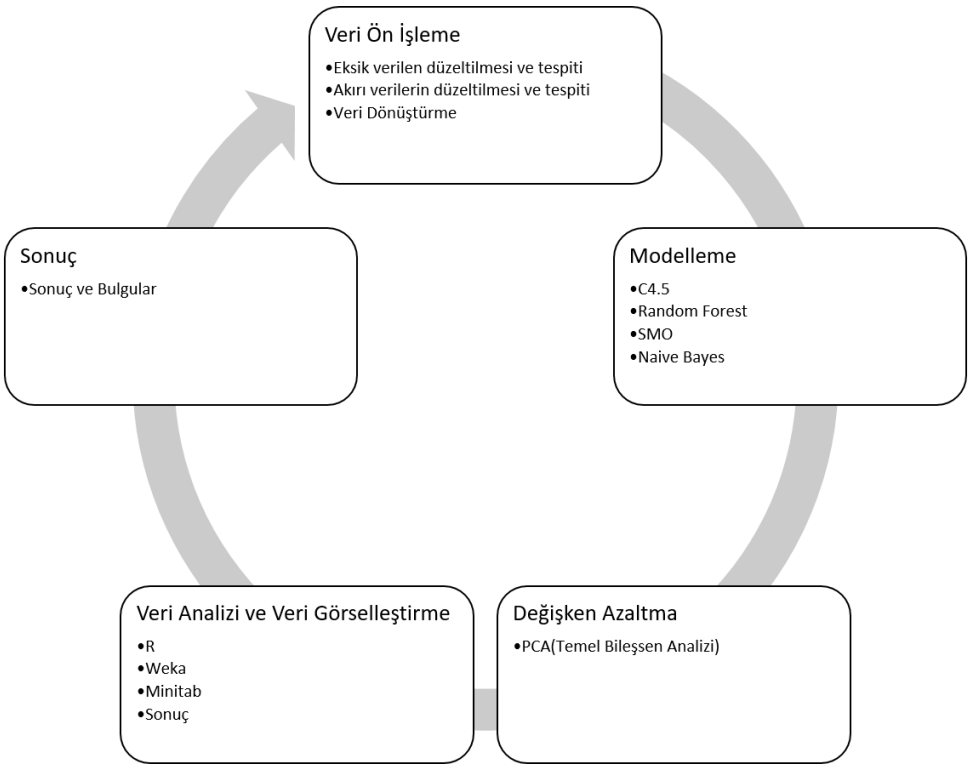
Bu tez çalışmasının amacı otomotiv sektöründe, kalite kontrol sürecinde, ölçüm veri setlerinin sınıflandırma temelli veri madenciliği yöntemleriyle hatalı parçaları en iyi tahmin eden sınıflandırma algoritmasını bulmaktır.

Bu çalışmanın ikinci bölümünde kalite kontrol, üçüncü bölümünde veri madenciliği dördüncü bölümünde veri madenciliği ve kalite kontrol süreci arasındaki ilişki, beşinci bölümünde yapılan çalışma ile kurulan modellere ve bu modellerin performanslarının yer aldığı sonuçlarla ilgili bilgiler yer almaktadır. Tez kapsamında yapılan uygulamaların değerlendirilmesi tartışma ve sonuç bölümünde yer almaktadır.

Veri madenciliği algoritmaları ile kurulacak olan modellerde veri madenciliği için çapraz endüstri standart süreci adımları kullanılmıştır [9].

Sınıflandırmaya dayalı veri madenciliği algoritmaları ile oluşturulan modeller kendi aralarında karşılaştırılarak en iyi sonucu veren modelin seçildiği kısım Bulgular bölümünde yer almaktadır. Çalışmanın son bölümü olan tartışma ve sonuç bölümünde

tezin genel bir değerlendirmesi ele alınmıştır.



Şekil 1.1. Tez çalışmasının genel işleyiş görünümü.

Tez kapsamında yapılan çalışmaların genel gösterimi aşağıda yer alan Şekil 1.1'de yer almaktadır. Veri ön işleme kısmında veri seti hazırlanırken yapılan işlemlere yer verilmiştir. Modelleme kısmında kurulan veri madenciliği modelleri anlatılmıştır. Değişken azalma, veri analizi, veri görselleştirme ve sonuç tezin diğer aşamalarıdır.

2. KALİTE KONTROL

Kalite (Qualites) Latince, “nasıl oluştuğu” anlamına gelen “qualis” kelimesinden türemiştir ve bir ürünün, istenen görevi daha iyi yapabilme (müşteri beklentilerini azami düzeyde sağlayan) veya her zaman aynı şekilde yapabilmesi (sürekli iyileştirme) için sahip olması gereken özellik olarak tanımlanmıştır [10].

Kalite ile ilgili değişik tanımlar mevcuttur. Bu tanımlardan bazıları aşağıda verilmiştir:

- Kalite, bir ürün ya da hizmetin belirlenen ya da olabilecek ihtiyaçları karşılama kabiliyetine dayalı özelliklerin toplamıdır (ISO).
- Kalite, bir mal ya da hizmetin belirli bir gerekliliği karşılayabilme yeteneklerini ortaya koyan karakteristiklerin tümüdür (ASQC).
- Kalite bir malın ya da hizmetin tüketicinin hizmetlerine uygunluk derecesidir.(EOQC).

Kontrol, mevcut sonuçlarla hedefleri ve amaçları kıyaslama ve gerekli olduğunda düzeltici önlemleri alma sürecidir. Herhangi bir kontrol sistemi üç bölümden oluşur:

1. Bir standart veya hedef
2. Bir başarıyı ölçme amacı
3. Ölçülen başarının standartlarla karşılaştırılması

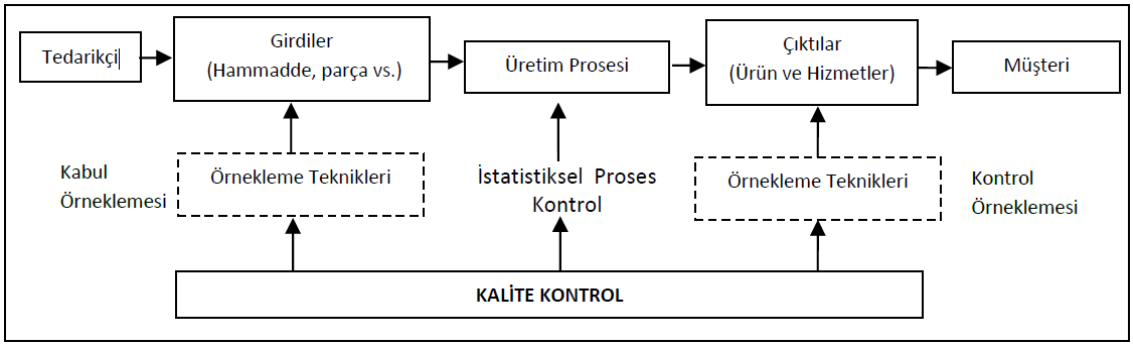
Kalite Kontrol;

1. Ürün veya hizmetin kullanım için uygun olmamasına neden olacak kusurlarını önleme, tespit etme ve düzeltme işlemidir.
2. Üretimi yapılan parça, ürün ve üniteden alınacak numunelerin incelenmesi suretiyle istenilen kalite seviyesine ulaşılması için yapılan işlemlere denir.
3. Bir ürünün tüketicisini tatmin etmesi ve onun beklentilerini en iyi biçimde karşılaması amacıyla üretimin her aşamasında sürdürülen denetim işlemleridir.
4. Üretimin planlanması aşamasında belirlenen kalite standartlarına üretim

işlemleri boyunca, öncesinde ve sonrasında ne ölçüde uyulduğunun incelenmesi ve gözlenmesidir.

Amaçları;

1. Satın alınan malzeme ve parçaların önceden belirlenmiş kalite standartlarını karşılamasını sağlamak.
2. İmalat prosesi süresince tasarım özelliklerine uygunluğu sürdürmek.
3. Nihai ürün veya hizmet için mümkün olan en yüksek kalite seviyesine ulaşmak.
    a. İmalattaki hurda ve yeniden işleme ile şikayet sayısı ve müşteriden geri dönüşleri azaltarak verimliliği artırma.
4. Kalite standartlarına ulaşılmadığında artan iç ve dış başarısızlık maliyetlerini azaltmak.



Şekil 2.1. Kalite kontrol süreci.

3. VERİ MADENCİLİĞİ

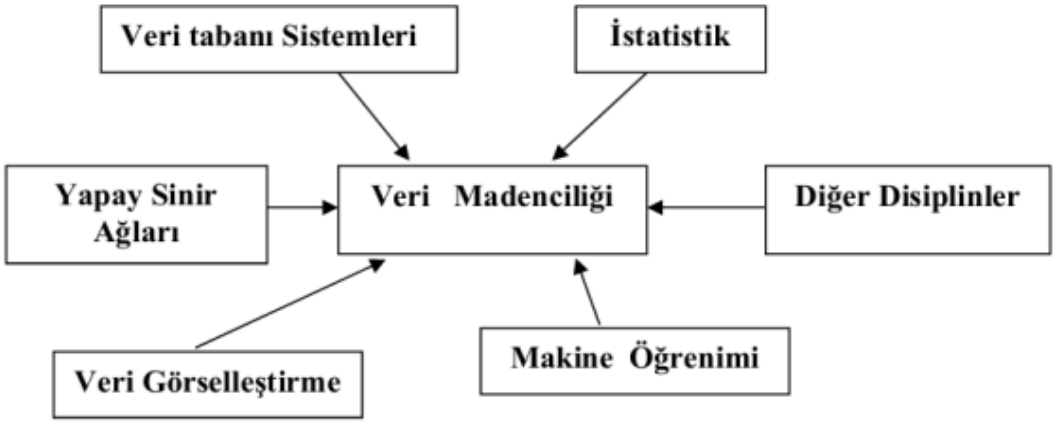
Veri madenciliği büyük miktarda bilginin depolanması ve faydalı bilginin ayrıştırılarak düzenlenmesi sürecidir. Başka bir tanımla; Veri tabanlarında bilgi keşfi, sıklıkla, büyük hacimde veri koleksiyonundan faydalı bilgiyi keşfetmeyi hedefleyen, Veri Madenciliği olarak adlandırılmaktadır [11]. Veri Madenciliği, pek çok analiz aracı kullanımıyla veri içerisinde örüntü ve ilişkileri keşfederek, bunları geçerli tahminler yapmak için kullanan bir süreçtir [12].

3.1. VERİ MADENCİLİĞİNİN TARİHSEL GELİŞİMİ

Veri madenciliğinin ilk çalışmaları 1950'lere dayanmaktadır ilk bilgisayarlarla beraber. Sayımlar için bilgisayarlar kullanılıyordu. 1960'larda veri koleksiyonu veri tabanı yaratımları başlamıştır. 1970'de ilişkisel veri modeli ve ilişkisel VTYS uygulamaları kullanılmıştır. 1980'lerde ilişkisel VTYS uygulamaları yaygınlaşmıştır. 1990'ra gelindiğinde veri boyutları büyümeye başlamıştır ve günlük işlemlerden derlenen büyük miktarda verinin nasıl değerlendirileceği araştırılmaya başlanmıştır. 1992'de veri madenciliği konusunda ilk yazılım gerçekleştirilmiştir. 2000'lere gelindiğinde Veri Ambarları ve Veri Madenciliği uygulamaları tamamen yaygınlaşmaya başlamıştır.

3.2. VERİ MADENCİLİĞİ ÇALIŞMA ALANLARI

Veri madenciliğinin günümüzde yaygın bir kullanım alanı vardır. Günümüzde tüm işletmeler sahip oldukları müşterilerin davranışlarını tahmin etmek istemektedirler. Veri madenciliği bu amaçla kullanılabilecek olan bir tekniktir. Bankacılık, pazarlama, sigortacılık, telekomünikasyon, borsa, tıp, endüstri, bilim ve mühendislik gibi alanlarda kullanılmaktadır. Şekil 3.1'de veri madenciliğinin kullanıldığı bazı alanlar gösterilmiştir [14].



Şekil 3.1.Veri madenciliğinin kullanıldığı alanlar.

3.2.1. Mühendislik Alanında Yapılan Çalışmalar

Mühendislik alanında veri madenciliği çalışmaları bulunmaktadır. Bu alanda yapılan çalışmalar mühendislik alanında kullanılan her türlü konuyu ele almaktadır. Örnek verecek olursak; 2007 yılında Kıyas Kayaalp tarafından yapılan bir çalışmada, veri madenciliği tekniği ile üç fazlı asenkron motordaki sargı spirleri arasında oluşabilecek kısa devre veya yalıtım bozuklukları ve motor milinde oluşabilecek mekanik dengesizlik hatalarının tespiti gerçekleştirilmiştir [13].

3.2.2. Tıp Alanında Yapılan Çalışmalar

Bu alanda da birçok veri madenciliği çalışması bulunmaktadır. Bu çalışmalarda hastaların hastalık, kan değerleri, şeker değerleri vb. sağlık verileri kullanılmaktadır. Tıp alanında yapılan çalışmalara örnek verilecek olursa; bir kişinin ailesinde olan bir hastalığın kendisinde ya da diğer aile üyelerinde olup olmadığına yönelik tahminsel çalışma, ölüm oranları ve salgın hastalıkların tahmin edilmesi gibi örnek çalışmalar yapılmaktadır. Bu çalışmaların ortak amacı olumlu sonuç elde ederek hastalık ihtimali olan kişileri bilinçlendirmek ve tedaviye yönlendirmektir. Harleen Kaur ve arkadaşları sınıflandırma yöntemlerinden karar ağacı ile model kurarak göğüs kanseri riskini tahmin etmeye çalışmışlardır. Bunun için hastaların yaş ve cinsiyet gibi verilerinden yararlanmışlardır. Günümüzde genetik mühendisleri bu çalışmaları geliştirmek amacıyla çalışmalar

gerçekleştirmektedirler [14].

3.2.3. Eğitim Alanında Yapılan Çalışmalar

Dünyada gerçekleştirilen veri madenciliği çalışma alanlarından biri de eğitim alanıdır. Eğitim alanında yapılan çalışmaların çoğu öğrenci başarısı üzerine yapılan analizleri içermektedir. Bu alanda gerçekleştirilen analiz uygulamalarının, sonraki nesiller için öngörü oluşturmak adına kullanılması, eğitim faaliyetlerine çok faydalı olacağı kanısına varılmıştır. Eğitim alanında yapılan birkaç veri madenciliği çalışmasını inceleyecek olursak;

Onur İnan tarafından hazırlanan çalışmada, hazırlık sınıfı, birinci sınıf ve mezun durumunda olan öğrenciler üzerinde, üniversite veri tabanındaki veriler kullanılarak; öğrencilerin başarılarını etkileyen etmenler, başarı düzeyleri, üniversiteyi kazanan öğrenci portföyleri ve mezun olamayan öğrencilerin okulu bitirmelerini etkileyen etmenler üzerinde çalışmalar gerçekleştirilmiş ve sonuçları yorumlanmıştır [15].

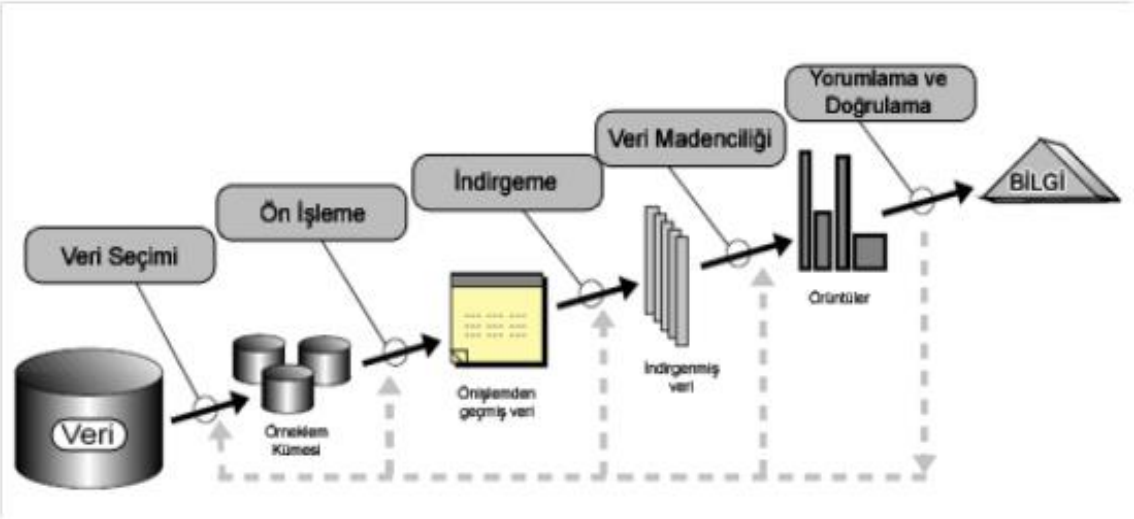
Serdar Çiftci tarafından hazırlanan çalışmada, uzaktan eğitime katılan öğrencilerin ders çalışma etkinliklerinin değerlendirilmesi için yapılan anketler ve log dosyaları karşılaştırılarak, sonuçların farklı olup olmadıkları üzerinde çalışılmıştır [16].

3.3. VERİ MADENCİLİĞİ SÜRECİ

Veri madenciliği çeşitli adımlardan meydana gelir ve bu adımlar sayesinde bir süreç olarak sonuca gider. Veri kaynağından alınan ham veriden algoritmaya giden yolda veri; temizlenir ve indirgenir. Daha sonrasında veri madenciliği metotları uygulanarak çeşitli bulgular ortaya çıkarılır. Bu bulgular yorumlanarak kararlar oluşturulur. Son aşamada görünen veririnin beklentileri en iyi şekilde karşılaması için bütün bu adımlar her noktasıyla ele alınmalı ve titiz bir şekilde uygulanmalıdır. Şekil 3.2'de görüldüğü üzere veri madenciliği aşamaları şu bölümlerden oluşmaktadır;

- Veri Seçimi
- Ön İşleme
- İndirgeme

- Veri Madenciliği
- Yorumlama ve Doğrulama



Şekil 3.2.Veri madenciliği süreci aşamaları.

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

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

İndirgeme; Veri üzerinden faydalı ve doğru sonuç elde etmek için kullanılacak verinin indirgenmesi gerekmektedir. Ancak büyük miktardaki verinin analizi zordur. Elde bulunan verinin büyük bir kısmı, her ne kadar ön işleme aşamasından geçmiş olsa da sonraki aşamalarda kullanılabilecek durumda değildir. Dolayısıyla kullanılabilecek duruma indirgenmesi lazımdır.

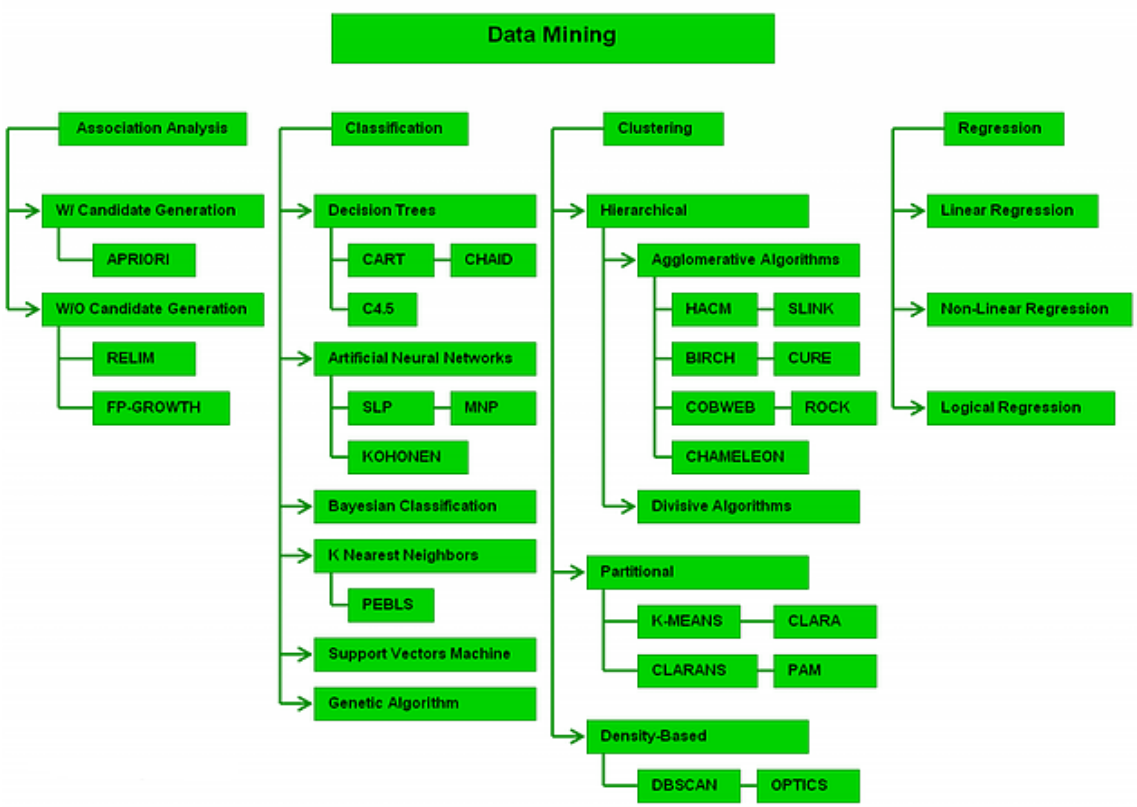
Veri Madenciliği; Veri seti hazırlanıncaya kadar olan aşama geçtikten sonraki aşamadır.

Veri bu aşamaya gelince doğru ve kullanılabilir bir formata gelmektedir. Çalışmanın amacına göre bu aşamada veri madenciliği yöntemlerinden biri veya birkaçı uygulanır. Gerektiği durumlarda farklı yöntemler birleştirilerek kullanılabilir.

Yorumlama ve Doğrulama; Modellerin oluşturulması tamamlandıktan sonraki adımdır. Veri üzerinde veri madenciliği uygulandıktan sonra alınan sonuçlar yorumlanır, karşılaştırılır ve çalışmanın doğru sonuca ulaşıp ulaşmadığı incelenir. Bu adımda genellikle farklı yöntemler uygulanmışsa onların karşılaştırması yapılır. Elde edilen sonuçlar yapılmış olan diğer çalışmaların sonuçlarıyla karşılaştırılıp doğrulanır [17].

3.4. VERİ MADENCİLİĞİ YÖNTEMLERİ

Veri madenciliği yöntemi, verinin veri madenciliği ile bilgiye nasıl dönüştürülebileceği yöntemleridir. Bilgi keşfinin hedeflenen sonuçlarına bağlı olarak çok farklı amaçlara sahip olabilirler. Bu yöntemlerin birçoğu istatiksel olmak üzere birçok algoritma barındırmaktadır. Kümeleme, Sınıflandırma, Regresyon ve Birliktelik Kuralları olarak dört ana grupta incelenmektedir. Her bir grup kendi içerisinde alt gruplara ve alt algoritmalara ayrılmaktadır. Şekil 3.3'te veri madenciliği yöntemlerinin hiyerarşisi gösterilmektedir.



Şekil 3.3.Veri yöntemleri.

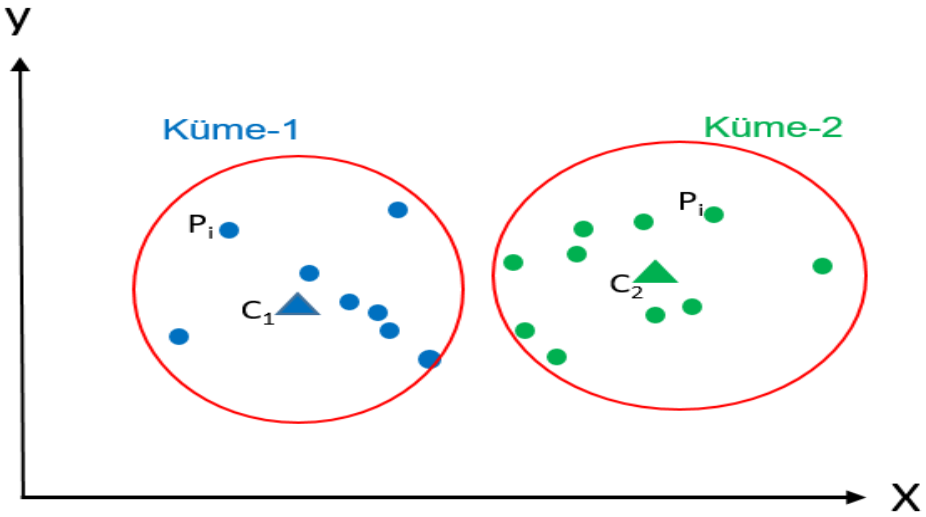
3.4.1. Sınıflandırma Yöntemi

Veriler arasındaki gizli bağlantı ve kuralları ortaya çıkarmak için kullanılan yöntemdir. Sıkça kullanılan yöntemlerden biridir. Bir niteliğin değerini diğer nitelikleri kullanarak belirlemek için önce verinin dağılımına göre bir model belirlenir. Daha sonra belirlenen bu model, başarı olanı belirlendikten sonra niteliğin bilinmeyen değerinin tahmin edebilmek için veya gelecekteki değerini bulmak için kullanılır. Karar ağaçları, Yapay Sinir Ağları, Bayes Sınıflandırıcılar, Bayes Ağları bu yöntemde en çok kullanılan modellerdir.

3.4.2. Kümeleme Yöntemi

Kümeleme, bir veri bilgilerini belirli yakınlıktaki kriterlere göre gruplara ayırma işlemidir. Bu grupların her birine “küme” adı verilir. Kümeleme işleminde küme içindeki elemanların benzerliği fazla, kümeler arası benzerlik ise az olmalıdır. Kümeleme veri madenciliği tekniklerinden tanımlayıcı modellere yani gözetimsiz sınıflandırmaya girer. Gözetimsiz sınıflamada amaç, başlangıçta verilen ve henüz sınıflandırılmamış bir küme, veriyi anlamlı alt kümeler oluşturacak şekilde öbeklemektir. Kümeleme işlemi tamamen

gelen verinin özelliklerine göre yapılır [18].



Şekil 3.4. Kümeleme yöntemi.

Şekil 3.4'te bir kümeleme yöntemi örneği verilmiştir. Koordinat düzleminde bulunan noktalar Küme-1 ve Küme-2 olmak üzere iki kümeye ayrılmıştır.

3.4.3. Birliktelik Kuralları Yöntemi

Olayların birlikte gerçekleşme durumlarını çözümleyen veri madenciliği yöntemlerine birliktelik kuralları adı verilmektedir. Bu yöntemler, birlikte olma kurallarını belirli olasılıklarla ortaya koyar. Birliktelik kuralları çoğunlukla büyük marketlerde kullanılmaktadır. Bu marketlerde müşterilerin satın aldıkları ürünler arasındaki kurallar oluşturularak bu kurallara göre ürünler yakın raflara yerleştirilmektedir.

3.5. VERİ MADENCİLİĞİ SÜRECİ ADIMLARI

Bu çalışmada veri madenciliği süreci adımları için CRISP modeli alınmıştır. CRISP modeli süreci aşağıdaki aşamalardan oluşmaktadır;

- İşi Anlamak
- Veriyi Anlama
- Veriyi Hazırlama
- Modelleme

- Değerlendirme

3.5.1. İşi Anlamak

CRISP sürecinin ilk adımı olup, üzerinde yapılan işin anlaşılmasıdır. Veri madenciliği çalışmasının yapılmaya çalışan işe ne katkı sağlayacağı netleştirilmeye çalışır. Bu adımda işi yapan kişinin bir problemi anlatması ve bu problemin veri madenciliği problemine indirgenmesi şeklinde oluşur. Problemin anlaşılması aşaması aşağıdaki adımlardan oluşmaktadır:

1. İşin amacının belirlenmesi
2. Mevcut durumun değerlendirilmesi
3. Veri Madenciliği amaçlarının belirlenmesi, hangi veri madenciliği yöntemleri kullanılacak karar verilmesi
4. Proje planının oluşturulması

3.5.2. Veriyi Anlamak

Sürecin ikinci adımı veriyi anlama adımıdır. Verinin anlaşılması aşamasında kullanılacak verinin nitelikleri belirlenir. Veriyi anlama aşağıdaki adımlardan oluşmaktadır:

1. Veri ilgili makine, cihaz, database vb. kaynaklardan toplanır ve bir araya getirilir
2. Veri üzerinde analiz yapılır
3. Veri üzerinde istatistik ve görselleştirme yaparak eldeki veri hakkında söylenecekler belirlenir.

3.5.3. Veriyi Hazırlama

Veriyi hazırlama adımı CRISP sürecinin en önemli adımlarından biridir. Çünkü bu aşamada ham veri seti üzerine operasyonlar yapılarak veri seti sadeleştirilip daha anlamlı daha iyi hale getirilir. Verilerden hangilerinin modelleme sürecinde kullanılacağı belirlenir. Daha sonra veri temizleme işlemi yapılır yani modelin oluşmasına engel olacak aykırı veriler temizlenir. Bu işlemden sonra aşağıda detaylı anlatılan operasyonlar veri seti üzerine uygulanır.

3.5.3.1. *Eksik Veriler*

Veri seti üzerinde her zaman veriler düzgün olmayabilir. Bazı durumlarda veri ölçümü gerçekleşmemiş, veri kaybolmuş vb. durumlar meydana gelmiş olabilir. Bu durumlar için eksik veri tamamlama adımları uygulanır. Bu adımları şu şekilde sıralayabiliriz [19]:

- Eksik nitelik değerleri olan veri kayıtlarını kullanma
- Eksik nitelik değerlerini elle doldur
- Eksik nitelik değerleri için global bir değişken kullan (Null, bilinmiyor,...)
- Eksik nitelik değerlerini o niteliğin ortalama değeri ile doldur
- Aynı sınıfa ait kayıtların nitelik değerlerinin ortalaması ile doldur
- Olasılığı en fazla olan nitelik değeriyle doldur

3.5.3.2. *Aykırı Veriler*

Aykırı veriler standart sapmanın dışında olan, normal şartlar altında olmaması gereken fakat ekstrem durumlardan dolayı meydana gelen verilerdir. Aykırı veriler veri seti üzene uygulanacak modellemeyi doğrudan etkileyeceğinden dolayı bu veriler düzeltilmelidir. Aykırı verileri düzeltme işlemi için adımları şu şekilde sıralayabiliriz:

- Bining, küçükten büyüğe veya büyükten küçüğe sıralanmış verileri düzeltmek için kullanılır
- Kümeleme, benzer gruplar aynı grup veya küme içinde yer alırken, aykırı değerler küme dışında kalacaktır.
- Regresyon, veriler regresyon ile verilere bir fonksiyon uydurularak düzeltilebilir. Uydurulan fonksiyona uymayan noktalar aykırı değerlerdir [20].

3.5.3.3. *Normalizasyon*

Veri setindeki niteliklerin değişken aralıkları çok geniş ise normalizasyon işlemi uygulanabilir. Min-max, z-scor, ondalık normalizasyon yöntemlerinden bir kaçıdır. Min-max yönteminde orijinal veriler yeni veri aralığına doğrusal dönüşüm ile dönüştürülür. Z-score yönteminde değişkenin her hangi bir y değeri, değişkenin ortalaması ve standart sapmasına bağlı olarak bilinen Z dönüşümü ile normalleştirilir. Ondalıkta ise, ele alınan değişkenin değerlerinin ondalık kısmı hareket ettirilerek normalleştirme yapılır [21].

- Min – Max Normalizasyon

  $v'=\frac{v-min_a}{max_{a-min_a}}(new\_max_A\text{-}new\_min_a)+new\_min_A$ (1)

- Z-score normalizasyon

  $v' = \frac{v-mean_A}{stand\_dev_A}$ (2)

- Ondalık normalizasyon

  $v' = \frac{v}{10^j}$ (3) j: MAX($|v'|$)<1 olacak şekildeki en küçük tam sayı

3.5.3.4. *Veri Dönüştürme*

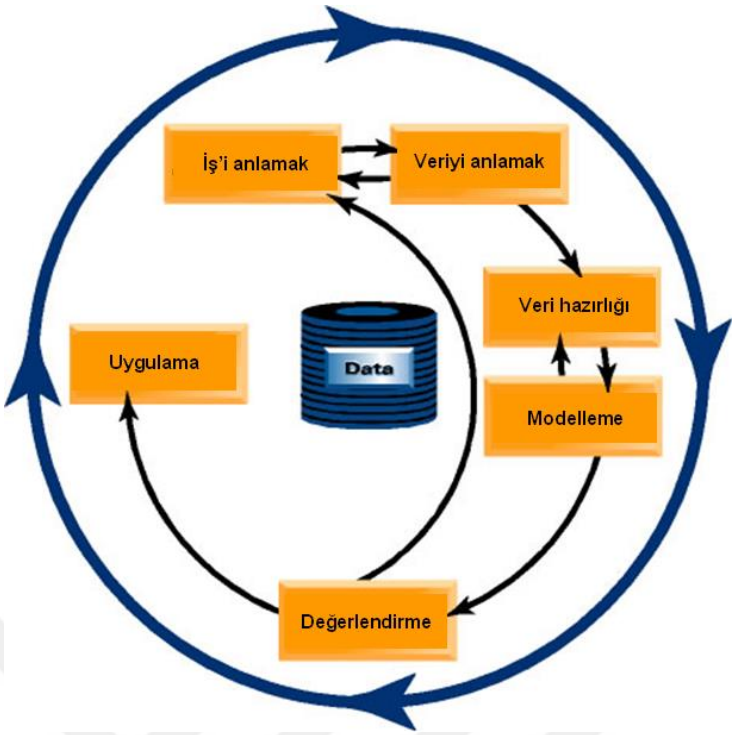
Bu adımda veri tipleri arasında dönüştürme yapılarak veri seti daha anlaşılır ve kullanılabilir hale getirilir. Örneğin cinsiyet niteliğindeki veri sütununda kayıtlar “Erkek” ve “Bayan” tipindeyken “0” ve “1” tipine dönüştürülür.

3.5.4. Modelleme

Bu adımda eldeki veriye ve istenen işe en uygun model seçimi yapılır. Modelleme adımları şu şekilde sıralanabilir:

- Modelleme sırasında kullanılacak yöntem oluşturulur, hangi algoritmanın kullanılacağı belirlenir
- Modelin kalitesini hangi test ile doğrulanacağı belirlenir. Örneğin sınıflandırma yapacaksak, verinin bir kısmını modelin geliştirmesi ve kalan kısmını da modelin testi için kullanarak test sırasında oluşturulan modelin karar verdiği sınıflandırmaların yüzde kaçında yanlış karar verdiğini belirleyebiliriz [22].
- Veri madenciliği amaçlarının oluşturulması
- Proje planının oluşturması

Şekil 3.5’te de görüldüğü gibi modelleme veri hazırlama adımıyla iki yönlü bir ilişki içindedir.



Şekil 3.5.CRISP süreci.

3.5.4.1. *C4.5 Algoritması*

C4.5 bir karar ağacı algoritmasıdır. ID3 algoritmasının bir üst seviyesidir. ID3 algoritmasında bazı eksiklikler vardı bunlar C4.5 algoritması ile çözülmüştür. C4.5 ile ID3 arasındaki en büyük fark normalizasyon yapılıyor olmasıdır. Ayrıca ID3 karar ağacından farklı olarak budama işlemi yapılır. C4.5 iki işlem adımı ile gerçekleştirilmektedir. Bunlardan ilki ağacı oluşturma işlemi ve diğeri ise budama işlemidir [23].

C4.5 algoritmasının çalışma mantığını inceleyecek olursak, birinci adım bilgi kazanımını hesaplamaktır.

Bilgi(M) = $-\sum_{i=1}^{k}((frekans(Si,M)/|M|)\,.\log_2(frekans(Si,M)/|M|))$

M: Herhangi bir misal

S: Sınıf

|M|: O sınıftaki tüm misallerin sayısı

Her nitelik için bilgi hesaplaması yapıldıktan sonra kazanım hesaplanmaya başlanır.

Kazanım(Özellik X) = Bilgi(P) - $\text{Bilgi}_x(P)$

Yani herhangi bir X özelliği için kazanım değeri, o özelliğin bağlı olduğu bütün parça ve

sadece o özelliği ilgilendiren parça arasındaki farka eşittir. Tüm kazanımlar hesaplandıktan sonra C4.5 ağacı en yüksek kazanıma sahip olan değeri alacaktır. Bu karar ağacının başlangıcı olacaktır. Daha sonra dalları oluşturmak için bu adımlar tekrar hesaplanır ve karar ağacı oluşturulur.

3.5.4.2. *Naive Bayes Algoritması*

Naive Bayes Sınıflandırıcı adını 17. yüzyılda yaşamış İngiliz matematikçi Thomas Bayes'ten alır. Naive Bayes sınıflandırıcı bağımsız varsayımlarla Bayes teoremini temel alan olasılıklı bir sınıflayıcıdır. Yalın tasarımına ve görünüşte basitleştirilmiş varsayımlara rağmen Naive Bayes sınıflandırıcı gerçek dünya durumlarında beklenenden çok daha iyi sonuçlar vermektedir [24]. Naive Bayes sınıflandırıcı ve tahmin edici algoritmadır. Bayes teoremi rastgele değişkenler için koşullu olasılıklar ile önsel olasılıklar arasındaki ilişkiyi verir.

Bayes Teoremi: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

P(A|B) : B olayı gerçekleştiği durumda A olayının meydana gelme olasılığı, P(B|A) ; A olayı gerçekleştiği durumda B olayının meydana gelme olasılığı, P(A) ve P(B) : A ve B olaylarının önsel olasılıklarıdır.

Naive Bayes sınıflandırma modelinde problem bir çok nitelikten ve bir sonuç değişkeninden meydana gelmektedir.

Naive Bayes Teoremi: $P(C|F_1,..,F_n) = \frac{P(C)p(F_1,..,F_n|C)}{p(F_1,..,F_n)}$

C verilen hedef ve F özelliklerimiz temsil eder. Naive bayes sınıflandırıcı basitçe bütün koşullu olasılıkların çarpımıdır.

3.5.4.3. *SMO Algoritması*

Optimizasyon Algoritması (SMO); SMO, esas itibariyle destek vektör kullanan bir algoritmadır. Çok terimli kernel kullanarak destek vektör sınıflandırıcıyı eğitmek için SMO Algoritmasını uygular. Bu uygulama global olarak bütün kayıp değerleri yenisiyle değiştirir ve nominal öznitelikleri ikili olanlara dönüştürür. Ayrıca bütün öznitelikleri önceden tanımlanmış değerlerle normalize eder [25].

3.5.4.4. *Random Forest Algoritması*

RF yönteminde, karar ormanını oluşturan karar ağaçları orijinal veri setinden bootstrap yöntemiyle seçilen farklı örneklerden oluşturulmaktadır. Her karar ağacında veri setindeki tüm değişkenlerden rastgele seçilen az sayıda değişken kullanılmaktadır. Her ağaç bir sınıf için oy vermektedir ve orman sınıflayıcısı bütün ağaçların verdiği oyları toplayarak bir sınıf için son tahminini yapmaktadır.

3.5.5. Değerlendirme

Sürecin son adımı değerlendirme adımıdır. Bu adımda ortaya çıkan sonuçlar veri madenciliği açısından incelenir.

4. MATERYAL VE YÖNTEM

4.1. WEKA, MINITAB VE R PROGRAMIYLA KALİTE TAHMİN UYGULAMASI

Bu bölümde veri madenciliği yöntemleri kullanılarak otomotiv sektöründe kalite kontrol sürecinde ürünlerin kalite tahminine yönelik bir uygulama yapılmıştır. Bu çalışma yapılırken CRISP modelinin adımları teker teker uygulanmıştır.

4.2. İŞİ ANLAMAK, PROBLEMİ TANIMLAMAK

Üretim sektöründe kalitenin önemli bir yeri bulunmaktadır, çünkü kalite iş gücü verimliliği ve müşteri memnuniyetinin temel noktasıdır. Üretim esnasında ortaya çıkan büyük ve karmaşık veriler nedeniyle toplam kaliteyi tahmin etmek zordur ayrıca hangi değişkenlerin hangi aralıkta olduğunda kalitede azalmaya neden olduğunu bulabilmek için istatistik kullanılmalıdır. Ancak istatiksel yöntemler karmaşık ve çok fazla zaman kaybettirmektedir. Ayrıca istatiksel yöntemler üretim bittikten sonra kaliteyi değerlendirebilir. İleriye yönelik bir tahminde bulunmaz. Bu çalışmanın ana amacı kaliteyi azaltan hatalı ürünlerin oluşup oluşmayacağı önceden tahmin etmektir. Bu sayede üretim esnasında ürüne müdahale edilebilir. Hatalı ürün üretimini tahmin etmek için sınıflandırma algoritmaları modeller oluşturuldu. Bu modeller belirli kriterlerle karşılaştırarak en iyi sonuç veren model bulunmuştur.

4.3. VERİYİ ANLAMAK

Üzerinde çalışılan veri seti Türkiye'deki bir otomotiv firmasının montaj hattındaki bir tezgâhtan toplanan ölçüm değerleridir. Verilerin doğruluğu %100 test edilmiştir. Ölçüm veri setine ait tüm değişkenler, gösterim biçimleri ve değerleri Çizelge 4.1'de gösterilmektedir. Ayrıca Çizelge 4.1'de değişkenlerin açıklamaları ve veri tipleri de verilmiştir.

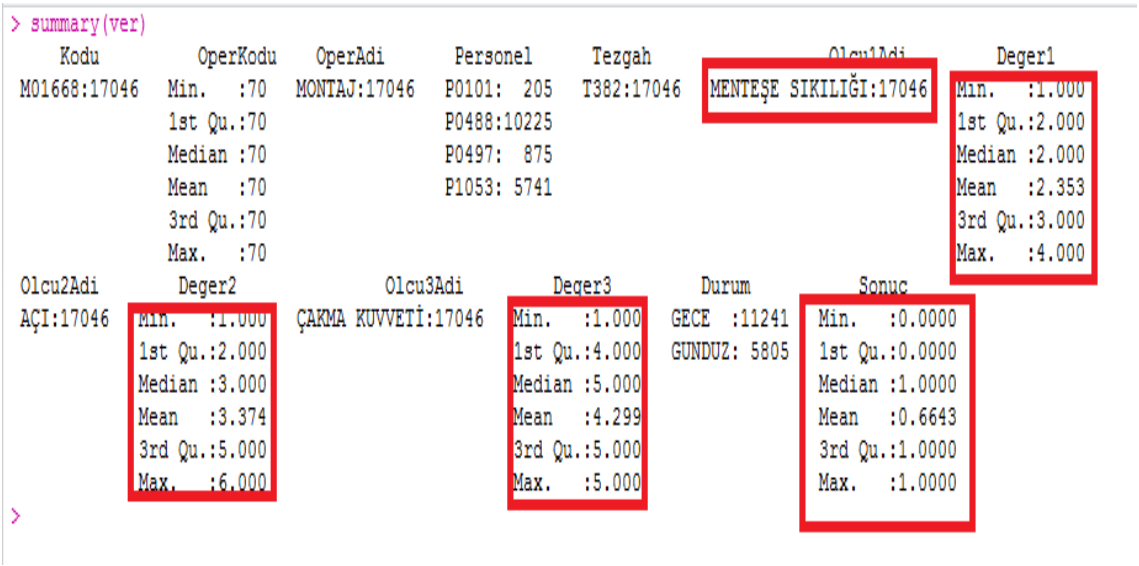
Çizelge 4.1. Ölçüm veri setine ilişkin tüm değişkenler,gösterim biçimleri ve tipleri

TAHMİN İÇİN KULLANILAN DEĞİŞKENLER			
NO	DEĞİŞKEN	AÇIKLAMASI	VERİ TİPİ
1	kodu	Ürünün Kodu	TEXT
2	adi	Ürünün Adı	TEXT
3	operKodu	Operasyonun Kodu	NUMERIC
4	operAdi	Yapılan Operasyonun Adı	TEXT
5	personel	Tezgâhta çalışan personelin kodu	TEXT
6	siraNo	Operasyon Sıra Numarası	NUMERIC
7	olcuAdi1	Bir numaralı ölçünün adı	TEXT
8	olcuDeger1	Bir numaralı ölçünün ölçülen değeri	NUMERIC
9	olcuTar1	Bir numaralı ölçünün tarihi	DATE
10	olcuSaat1	Bir numaralı ölçünün saati	DATE
11	olcuAdi2	İki numaralı ölçünün adı	TEXT
12	olcuDeger2	İki numaralı ölçünün ölçülen değeri	NUMERIC
13	olcuTar2	İki numaralı ölçünün tarihi	DATE
14	olcu2Saat	İki numaralı ölçünün saati	DATE
15	olcuAdi3	Üç numaralı ölçünün adı	TEXT
16	olcuDeger3	Üç numaralı ölçünün ölçülen değeri	NUMERIC

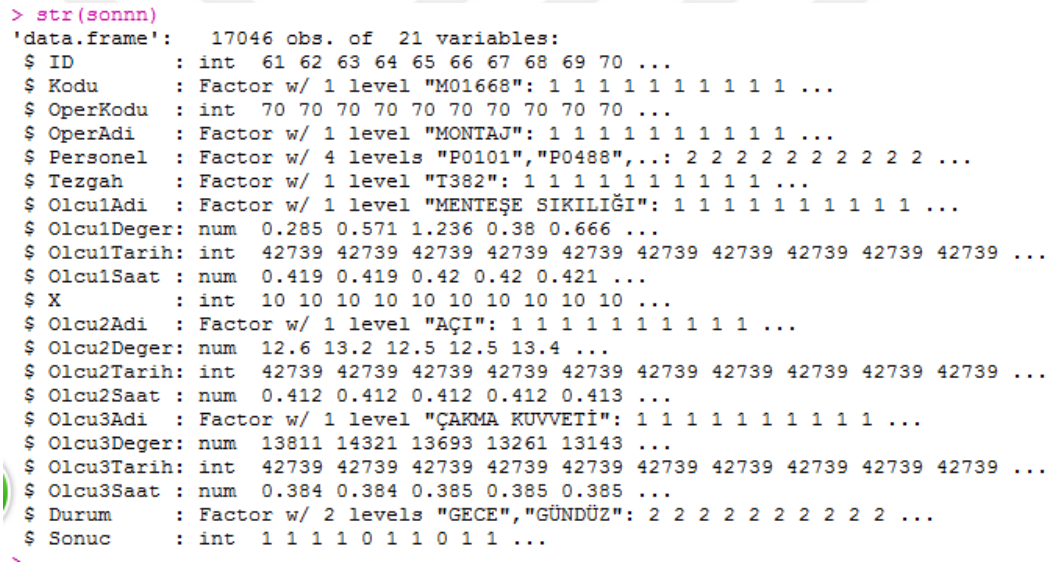
Çizelge 4.1 (devam). Ölçüm veri setine ilişkin tüm değişkenler,gösterim biçimleri ve tipleri

17	olcuTar3	Üç numaralı ölçünün tarihi	DATE
18	olcu3Saat	Üç numaralı ölçünün saati	DATE
19-22	nominal_value	Ölçülerin değerinin de nominal değerleri	NUMERIC
23-26	nominal_poz	Pozitif yönde aralık	NUMERIC
27-30	nominal_neg	Negatif yönede aralık	NUMERIC
31-34	temperature	Ölçüm sırasında ortam sıcaklığı	NUMERIC
35	tork	Motorun Tork Kuvveti	NUMERIC
36	location	Ürüne kuvvet uygulayan kısmın lokasyon değeri	NUMERIC
37	result	Ürünün hatalı olup olamadığının bilgisi	İKİLİ

Çizelge 4.1 incelendiğinde ölçüm veri setinin değerlerinin text, numeric, date ve ikili değerlere ayrıldığı görülmektedir. Toplam 37 adet nitelik bulunmaktadır. Şekil 4.1 de veri seti özet bilgisine bakıldığında ölçüm değerlerinin minimum, maksimum, ortalama ve medyan değerleri verilmiştir. Text değişkenlerinin aldığı değerler ait frekans değerleri de verilmiştir. Eksik veriler daha önceden tamamlandığından dolayı eksik verilerle ilgili bilgi verilmemiştir.



Şekil 4.1.Veri özeti.

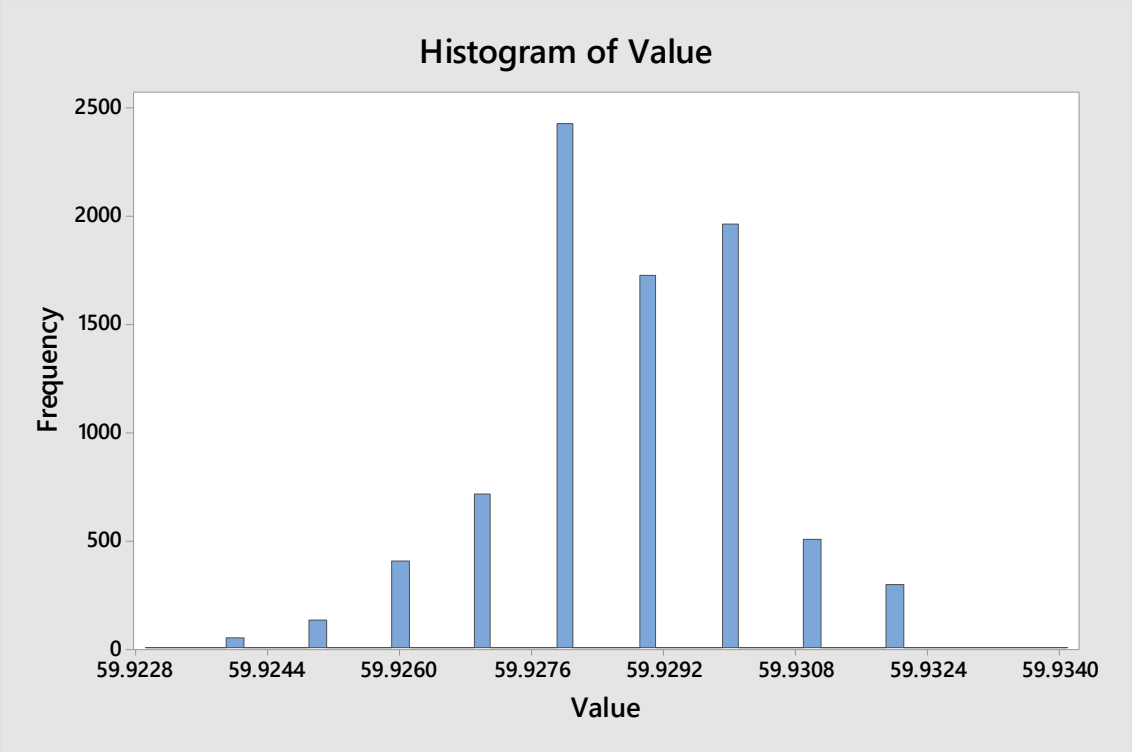


Şekil 4.2. Veri seti gösterim biçimleri, türleri

Şekil 4.1 ve Şekil 4.2 R programında veri seti analizi sonucunda hazırlanmıştır. Şekil 4.2'de veri indirgeme, normalizyon ve eksik veriler düzeltilmiş veri setinin analizi gözükmektedir.

Bu çalışma kapsamında üretim verileri görselleştirme tekniklerinden elde edilen kalite ölçüm veri setine histogram yöntemi uygulanmıştır. Bu uygulama Minitab üzerinde gerçekleştirildi. Şekil 4.3'de gösterilen histogram, talaşlı imalat bölümündeki işleme operasyonu esnasında ölçülen çap ölçüm değerleridir. Bu ölçümün nominal değeri 59.923, nominal artı değeri 0.012 ve nominal eksi değeri 0'dır. Bu durumda, Alt Kontrol

Limiti (AKL) 59.923 ve Üst Kontrol Limiti (UCL) 59.937'dir. Histogram, sınıf aralığı ile orantılıdır ve yükseklik, tekrarların sayısını belirten sınıf frekanslarıyla orantılıdır. Elde edilen en büyük ve en küçük değerler arasındaki fark, değişim genişliği olarak bulunur. Varyasyon aralığı, sınıf aralığını belirlemek için sınıf sayısına bölünür.



Şekil 4.3.İşleme Operasyonundaki çap değer ölçüm histogramı.

4.4. VERİYİ HAZIRLAMAK

Verilerin analizlere hazırlanmasında bölüm 3.3' teki veri madenciliği süreçleri ve 3.5'de anlatılan veri madenciliği süreci adımları dikkate alınmıştır. Bu tez çalışmasında takip edilen veri madenciliği süreçlerinde veri temizleme, veri dönüştürme ve normalizasyon kullanılmıştır. Veri ön işleme adımından sonra modelleme adımına geçilmiştir.

4.4.1. Veri Temizleme

Veri temizle veri setine uygulanacak model için önemli bir yere sahiptir. Çünkü veri seti ne kadar düzgünse uygulanacak model o derece iyi performans gösterir. Bu bölümde eksik verilerin temizlenmesi ve tamamlanması, aykırı verilerin tespiti ve çözümlenmesi incelenecektir.

Bu çalışmada kullanılacak veri seti veri tabanından alındı. Bu nedenle veri tabanından veri setini alırken veri temizle ve diğer işlemler veri seti oluşturulurken gerçekleştirildi. Ölçü değerleri olan değişkenler üzerinde ki eksik kayıtlar veri seti hazırlanırken çıkartıldı. Yine aynı şekilde date veri tipinde olan değişkenlerde "Null" veri tipi varsa bu satırlar da veri setine dâhil edilmedi. Operasyon Kodu, Ürün Kodu vb. değişkenlerdeki eksikliklerde eksik veriler yerine olası en muhtemel değişkenler atandı. Sıcaklık değişkenindeki eksik verilerde sütundaki verilerin ortalaması eksik verilerin yerine kullanıldı. Eksik verilerin dışında gürültü veriler üzerinde de veri ön işleme operasyonları uygulandı. Veri setindeki numeric olan nitelikteki veriler üzerinde gürültüyü azaltmak için demetleme yöntemi kullanıldı.

4.5. MODELLEME

Bu tez çalışmasında otomotiv sektöründeki otomobil parçalarının hatalı olup olmayacağını tahmin etmeye çalışan veri madenciliği yöntemlerinde sınıflandırma algoritmaları ile modelleme uygulamaları yapılmıştır. Bu sınıflandırma algoritmaları C4.5, SMO, Random Forest ve Naive Bayes algoritmalarıdır.

Bu algoritmalar oluşturulan veri setine uygulanarak hangi modelin daha iyi olduğuna karar verilmeye çalışılmıştır. Bu karar verme aşamasında hold-out ve çapraz geçerleme ve performans değerlendirme analiz yöntemleri kullanılarak çıkan sonuçlar iki faklı yöntemle değerlendirilmiştir. İlk yöntem olan hold-out için %40-%60, %25-%75, %20-%80 ayrım oranlarına sahip test ve eğitim veri seti ayrımı yapılarak değerlendirilmiştir. İkinci değerlendirme yöntemi olan çapraz geçerleme ile 5-kat ve 10-kat çapraz geçerleme yapılmıştır. Bu şekilde iki farklı analiz yöntemi kullanarak bulduğumuz sonucun doğruluğunun ispatlamasını da yapmış olduk.

Bu modelleme süreci weka, R ve minitab programları aracılığıyla gerçekleşmiştir. Weka veri madenciliği algoritmalarının neredeyse tamamını üzerinde bulunduran ve bu algoritmaların kullanımını sağlayan, ayrıca veri görselleştirme, veri analizi, iş zekâsı uygulamaları gibi benzer özellikleri üzerinde bulunduran modüler bir programdır. Bu tez çalışmasındaki veri setine algoritma uygulanması, kuralların ve karar ağaçlarının oluşturulması gibi işlemler bu program üzerinde gerçekleştirilmiştir. R istatistikçi ve matematikçilerin yoğun bir şekilde tercih ettiği bir istatistik ve analiz programıdır. Tez

çalışmasında kullanılan veri seti üzerinde analiz işlemleri R paketleri ile gerçekleştirilmiştir. Bu çalışma kapsamında ilaveten kullanılan bir diğer program da minitab ‘dır. Minitab da yine R gibi bir istatistik programıdır. Bu çalışmada minitab ile kalitenin daha iyi yorumlanabilmesi için bir veri görselleştirme uygulamasına yer verilmiştir.

5. BULGULAR

Bu bölümde, C4.5, Random Forest, SMO ve Naive Bayes algoritmalarının üretim kalite ölçüm veri seti üzerinde uygulanması ve elde edilen modellerin karşılaştırma sonuçlarına yer verilmiştir. Her model için hedef değişken, performans değerlendirme ve model seçim yöntemi ve kullanılan programlar ve yapılan işlemler sırasıyla Çizelge 5.1, Çizelge 5.2, Çizelge 5.3 ve Çizelge 5.4'te verilmiştir.

5.1. C4.5 ALGORİTMASI İLE MODEL KURMA

C4.5 modeli oluşturulurken holt-out ve çapraz gerçerleme ve performans değerlendirme yöntemleri kullanılmıştır.

Çizelge 5.1'de görüldüğü gibi hold-out yöntemde eğitim ve test veri setinin %60-%40, %75-%25, %80-%20 şeklinde sırayla ayrımı yapılmıştır. k-kat çapraz geçerlemeden de 5-kat çapraz geçerleme ve 10-kat çapraz geçerleme kullanılmıştır. Modellerin karşılaştırılması performans karşılaştırılması bölümünde verilmiştir. Ayrıca hold-out ve çapraz geçerleme için kodlar ekler bölümüne eklenmiştir.

Çizelge 5.1. C4.5 algoritma model özeti.

Hedef Değişken	Result (E/H-Evet/Hayır)
Performans Değerlendirme ve Model Seçim Yöntemi	• 5-kat çapraz geçerleme ve 10-kat çapraz geçerleme • %60-%40, %75-%25, %80-%20 oranlarında hold-out
WEKA ile yapılan işlemler	• Veri setinin. arff formatına dönüştürülmesi • Veri setinin programa upload edilmesi • C4.5 algoritmasının uygulanması
R ile Yapılan İşlemler	• Şekil 4.1 ve 4.2'deki veri analizlerinin yapılması
Minitab ile Yapılan İşlemler	• Histogram Çıkartımı

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

Deger2 <= 1: 0 (2153.0)
Deger2 > 1
|   Deger3 <= 2: 0 (1163.0)
|   Deger3 > 2
|   |   Deger2 <= 5: 1 (12141.0/1689.0)
|   |   Deger2 > 5
|   |   |   Deger3 <= 4: 1 (550.0/207.0)
|   |   |   Deger3 > 4
|   |   |   |   Durum = GECE
|   |   |   |   |   Personel = P0101: 0 (3.0/1.0)
|   |   |   |   |   Personel = P0488: 0 (494.0/235.0)
|   |   |   |   |   Personel = P0497: 1 (9.0/4.0)
|   |   |   |   |   Personel = P1053: 1 (185.0/86.0)
|   |   |   |   Durum = GÃœNDÃœZ
|   |   |   |   |   Personel = P0101: 1 (0.0)
|   |   |   |   |   Personel = P0488: 1 (244.0/104.0)
|   |   |   |   |   Personel = P0497: 0 (46.0/19.0)
|   |   |   |   |   Personel = P1053: 1 (58.0/28.0)

Number of Leaves  :     12

Size of the tree :      19
```

Şekil 5.1.C4.5 Algoritması veri seti kuralları.

Şekil 5.1'de C4.5 algoritması veri setine uygulandıktan sonra oluşan kurallar gösterilmektedir. Bu kuralları açıklayacak olursak;

Kural 1: Eğer Değer2 "1" e eşit veya küçük ise parça hatalıdır. Result durumu "0" olur.

Kural 2: Eğer Değer2 büyüktür "1" ve Değer3 "2" den küçük veya eşit ise parça hatalıdır. Result durumu "0" olur.

Kural 3: Eğer Değer3 büyüktür "2" ve Değer2 küçük eşit "5" ise parça düzgündür. Result durumu "1" olur.

Kural 4: Değer2 büyüktür "5" ve Değer3 küçük veya eşit "4" ise parça düzgündür. Result durumu "1" olur.

Kural 5: Değer3 büyüktür "4" ve Durum eşittir "GECE" ve Personel eşittir "P0101" ise parça hatalıdır. Result durumu "0" olur.

Kural 6: Değer3 büyüktür "4" ve Durum eşittir "GECE" ve Personel eşittir "P0488" ise parça hatalıdır. Result durumu "0" olur.

Kural 7: Değer3 büyüktür “4” ve Durum eşittir “GECE” ve Personel eşittir “P0497” ise parça düzgündür. Result durumu “1” olur.

Kural 8: Değer3 büyüktür “4” ve Durum eşittir “GECE” ve Personel eşittir “P1053” ise parça düzgündür. Result durumu “1” olur.

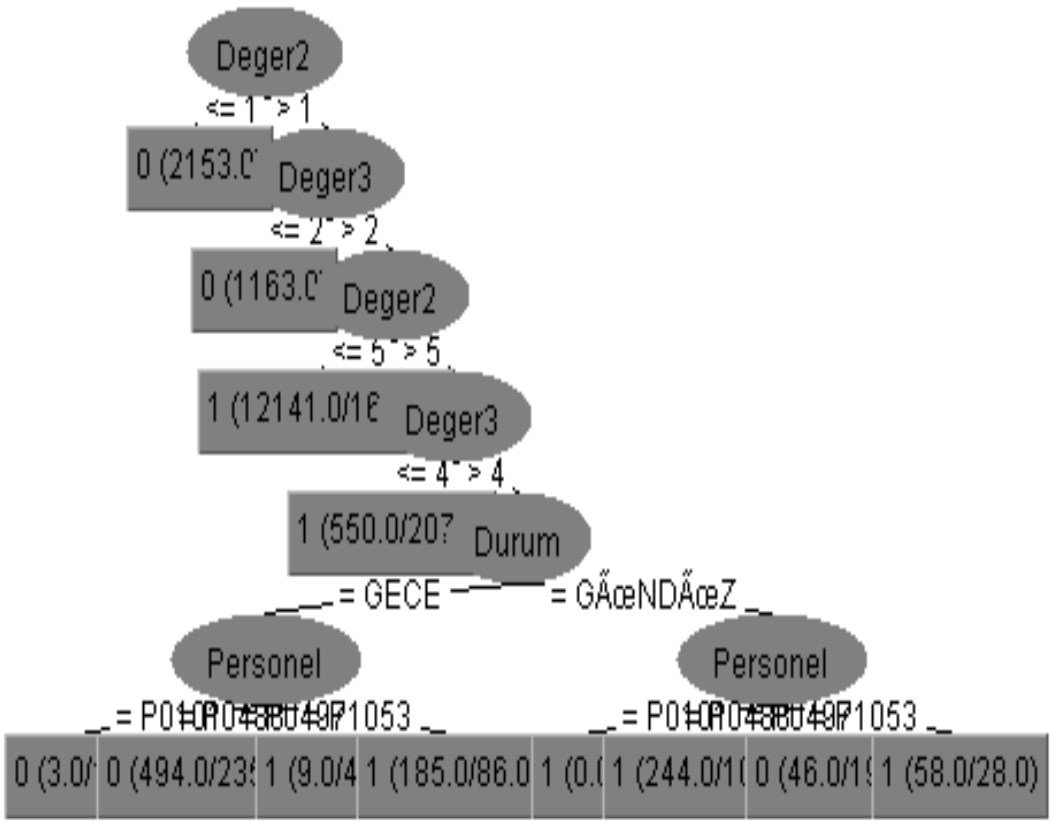
Kural 9: Değer3 büyüktür “4” ve Durum eşittir “GÜNDÜZ” ve Personel eşittir “P0101” ise parça düzgündür. Result durumu “1” olur.

Kural 10: Değer3 büyüktür “4” ve Durum eşittir “GÜNDÜZ” ve Personel eşittir “P0488” ise parça düzgündür. Result durumu “1” olur.

Kural 11: Değer3 büyüktür “4” ve Durum eşittir “GÜNDÜZ” ve Personel eşittir “P0497” ise parça hatalıdır. Result durumu “0” olur.

Kural 12: Değer3 büyüktür “4” ve Durum eşittir “GÜNDÜZ” ve Personel eşittir “P1053” ise parça düzgündür. Result durumu “1” olur.

Şekil 5.2’de C4.5 algoritmasından oluşan karar ağacı gösterilmektedir. Ağacın yapısı Şekil 5.1’deki kurallara göre oluşmaktadır.



Şekil 5.2.C4.5 Algoritması karar ağacı.

5.2. RANDOM FOREST ALGORİTMASI İLE MODEL KURMA

Random Forest yani rastgele orman modeli oluşturulurken sınıflandırma işlemi esnasında birden fazla karar ağacı oluşturularak sınıflandırma değerini yükseltmek hedeflendi. Daha sonra elde edilen sonuçlarda holt-out ve çapraz gerçerleme ve performans değerlendirme yöntemleri kullanılmıştır.

Çizelge 5.2'de görüldüğü gibi hold-out yöntemde eğitim ve test veri setinin %60-%40, %75-%25, %80-%20 şeklinde sırayla ayrımı yapılmıştır. k-kat çapraz geçerlemeden de 5-kat çapraz geçerleme ve 10-kat çapraz geçerleme kullanılmıştır. Modellerin karşılaştırılması performans karşılaştırılması bölümünde verilmiştir. Ayrıca hold-out ve çapraz geçerleme için kodlar ekler bölümüne eklenmiştir.

Çizelge 5.2. Random Forest algoritma model özeti.

Hedef Değişken	Hatasız (E/H-Evet/Hayır)
Performans Değerlendirme ve Model Seçim Yöntemi	• 5-kat çapraz geçerleme ve 10-kat çapraz geçerleme • %60-%40, %75-%25, %80-%20 oranlarında hold-out
WEKA ile yapılan işlemler	• Veri setinin. arff formatına dönüştürülmesi • Veri setinin programa upload edilmesi • Random Forest algoritmasının uygulanması
R ile Yapılan İşlemler	• Şekil 4.1 ve 4.2'deki veri analizlerinin yapılması
Minitab ile Yapılan İşlemler	• Histogram Çıkartımı

5.3. SMO ALGORİTMASI İLE MODEL KURMA

Ardışık minimal optimizasyon (SMO) modeli oluşturulurken holt-out ve çapraz geçerleme ve performans değerlendirme yöntemleri kullanılmıştır.

Çizelge 5.3'de görüldüğü gibi hold-out yöntemde eğitim ve test veri setinin %60-%40, %75-%25, %80-%20 şeklinde sırayla ayrımı yapılmıştır. k-kat çapraz geçerlemeden de 5-kat çapraz geçerleme ve 10-kat çapraz geçerleme kullanılmıştır. Modellerin karşılaştırılması performans karşılaştırılması bölümünde verilmiştir. Ayrıca hold-out ve çapraz geçerleme için kodlar ekler bölümüne eklenmiştir.

Çizelge 5.3. SMO algoritma model özeti.

Hedef Değişken	Hatasız (E/H-Evet/Hayır)
Performans Değerlendirme ve Model Seçim Yöntemi	• 5-kat çapraz geçerleme ve 10-kat çapraz geçerleme • %60-%40, %75-%25, %80-%20 oranlarında hold-out
WEKA ile yapılan işlemler	• Veri setinin. arff formatına dönüştürülmesi • Veri setinin programa upload edilmesi • SMO algoritmasının uygulanması
R ile Yapılan İşlemler	• Şekil 4.1 ve 4.2'deki veri analizlerinin yapılması
Minitab ile Yapılan İşlemler	• Histogram Çıkartımı

5.4. BAYES ALGORİTMASI İLE MODEL KURMA

Naive Bayes modeli oluşturulurken holt-out ve çapraz geçerleme ve performans değerlendirme yöntemleri kullanılmıştır.

Çizelge 5.4'de görüldüğü gibi hold-out yöntemde eğitim ve test veri setinin %60-%40, %75-%25, %80-%20 şeklinde sırayla ayrımı yapılmıştır. k-kat çapraz geçerlemeden de 5-kat çapraz geçerleme ve 10-kat çapraz geçerleme kullanılmıştır. Modellerin karşılaştırılması performans karşılaştırılması bölümünde verilmiştir. Ayrıca hold-out ve çapraz geçerleme için kodlar ekler bölümüne eklenmiştir.

Çizelge 5.4. Bayes algoritma model özeti.

Hedef Değişken	Hatasız (E/H-Evet/Hayır)
Performans Değerlendirme ve Model Seçim Yöntemi	• 5-kat çapraz geçerleme ve 10-kat çapraz geçerleme • %60-%40, %75-%25, %80-%20 oranlarında hold-out
WEKA ile yapılan işlemler	• Veri setinin. arff formatına dönüştürülmesi • Veri setinin programa upload edilmesi • Naive Bayes algoritmasının uygulanması
R ile Yapılan İşlemler	• Şekil 4.1 ve 4.2'deki veri analizlerinin yapılması
Minitab ile Yapılan İşlemler	• Histogram Çıkartımı

5.5. MODEL PERFORMANS KARŞILAŞTIRILMASI

Bu bölümde oluşturan tüm modellerden alınan sonuçlar analiz edilip tablolar halinde gösterilerek performansları karşılaştırılmıştır. Birinci tabloda veri seti uygulanan modellerin 5 kat çapraz geçerleme, 10 kat çapraz geçerleme ile bulunan sonuçları verilmiştir. İkinci tabloda ise %60-%40, %75-%25, %80-%20 katsayılarında holt-out yöntemi uygulanarak bulunan sonuçlar verilmektedir.

5.5.1. Çapraz Geçerleme Performans Değerlendirmesi ve Model Seçimi ile Elde Edilen Bulgular

Bu tez çalışmasında, dört adet veri madenciliği sınıflandırma model oluşturulmuştur. Bu modellerin performans değerlerinin daha iyi anlaşılabilmesi için 5 kat çapraz geçerleme ve 10 kat çapraz geçerleme veri seti üzerinde uygulanmıştır. Tablo 5.5'de veri setine 5 kat çapraz geçerleme ve 10 kat çapraz geçerleme uygulanarak bulunan sonuçlar verilmiştir. Bu sonuçlara göre C4.5 modeli en iyi performansı sağlayan model olmuştur. Ayrıca Random Forest modeli C4.5 modeline çok yakın sonuçlar vermiş olsa da süre performansı olarak çok geride kalmıştır. Detaylı performans özellikleri ve verileri "Ekler" kısmına eklenmiştir.

Çizelge 5.5.5-Kat ve 10-Kat çapraz geçerleme performans değerlendirme sonuçları.

	5 kat çapraz geçerleme				10 kat çapraz geçerleme			
FEATURE	C4.5	N.Bayes	SMO	Rand.F	C4.5	N.Bayes	SMO	Rand.F
Doğruluk	0.8576	0.7957	0.6938	0.8546	0.8581	0.7989	0.6938	0.8564
Hata	0.1424	0.2043	0.3062	0.1454	0.1419	0.2011	0.3062	0.1436
Precision	0.871	0.805	0.738	0.862	0.873	0.808	0.738	0.862
Recall	0.858	0.796	0.694	0.855	0.858	0.799	0.694	0.855
F-measure	0.848	0.778	0.604	0.847	0.848	0.782	0.604	0.847
Roc Area	0.828	0.761	0.549	0.891	0.827	0.762	0.549	0.891
Run time	0.42	0.8	50.31	10.73	5.59	0.08	56.34	12.05

5.5.2. Hold-Out Performans Değerlendirmesi ve Model Seçimi ile Elde Edilen Bulgular

Bu tez çalışmasında, hold-out performans değerlendirme yöntemiyle modelleri

değerlendirirken sırasıyla test ve eğitim kümeleri %40-%60, %25-%75, %20-%80 ayrımları ile karşılaştırılmış ve sonuçlar Tablo 5.6'te verilmiştir. Bu sonuçlara göre C4.5 modeli hold-out yönteminde de en iyi performansı sağlayan model olmuştur. Ayrıca Random Forest modeli yine C4.5 modeline çok yakın sonuçlar vermiş olsa da süre performansı olarak çok geride kalmıştır. Detaylı performans özellikleri ve verileri "Ekler" kısmına eklenmiştir.

Çizelge 5.6.Hold-Out performans değerlendirme sonuçları.

	DOĞRULUK			HATA		
%T-%E	40-60	25-75	20-80	40-60	25-75	20-80
C4.5	0.8580	0.8556	0.8524	0.142	0.1444	0.1476
N.BAYES	0.7855	0.7793	0.7688	0.2145	0.2207	0.2312
SMO	0.7224	0.7169	0.7160	0.2776	0.2831	0.284
RAND. F	0.8580	0.8549	0.8503	0.142	0.1451	0.1497

6. TARTIŞMA VE SONUÇ

Bu tez çalışması boyunca kalite kontrol, veri madenciliği, kalite kontrol ile veri madenciliği arasındaki ilişkiler, veri madenciliği yöntemleri ve modelleri konuları anlatılmıştır. Kalite kontrol ve üretimde veri madenciliği konuları üzerine literatür taraması yapılmıştır. Yapılan çalışmaların çoğunda sınıflandırma modeli kullanılmıştır. Bu nedenden dolayı tez çalışmasında da veri madenciliği yöntemlerinden sınıflandırma modelindeki algoritmalar kullanılmıştır. Çalışmaya ek olarak veri madenciliği yöntemlerinden veri görselleştirme yöntemiyle de ek bir uygulama yapılmıştır.

Bölüm 5.1'de C4.5, 5.2'de Random Forest, 5.3'de Sequential Minimal Optimisation (SMO), 5.4'de Naive Bayes modellerinin nasıl oluşturulduğu verilmiştir. 5.5.1 de ise bu oluşturulan modellerin 4-kat çapraz geçerleme ve 5 kat çapraz geçerleme performans değerlendirmelerinin detaylı sonuçları karşılaştırılmaları verilmiştir. Bölüm 5.5.2 de hold-out yöntemi kullanılarak performans değerlendirme, sonuç ve karşılaştırmalar tablo halinde verilmiştir.

Hold-out yönteminde veri seti %20-%80, %25-%75, % 40-%60 sırasıyla test ve eğitim veri setine ayrılmış C4.5 yaklaşık %86'lık doğruluk oranı ile yine en iyi performansı gösteren model olmuştur. Random Forest ikinci, Naive Bayes üçüncü ve SMO dördüncü en iyi performansı gösteren algoritma olmuştur. C4.5 ve Random Forest modellerinin çapraz geçerleme yöntemiyle çıkan performans sonuçları ile hold-out yöntemiyle çıkan sonuçları karşılaştırıldığında performans olarak büyük değişiklik olmadığı görülmüştür. Naive Bayes modelinin hold-out yöntemiye daha kötü performasn verirken, SMO modelinin daha iyi performans verdiği tespit edilmiştir.

Bölüm 6'da veri madenciliği yöntemlerinden olan veri görselleştirme yöntemiyle bir histogram oluşturulmuş ve bu histogram kalite değişimi ve performansı açısından incelenmiştir. Bu çalışmada hedef yüzlerce rakamın yer aldığı tablolarla uğraşmamak ve veri madenciliği tekniklerine girmeden grafikler ile elimizde bulunan veri seti hakkında bilgi sahibi olabilmeyi hatta yorum yapmayı sağlayan veri görselleştirme yönteminin uygulanmasıdır. Ayrıca histogramın çıkarıldığı sektör ve veri seti hakkında tecrübesi

bulunan herkes, elinde bulunan veri setine ilişkin buna benzer çalışmalar yapabilir ve elde ettiği sonuçlarla ilgili yorum yapabilme imkânına sahip olabilir. Bu şekilde herhangi bir veri setine uygun veri görselleştirme tekniği uygulandığında veri seti hakkında yorum yapmak ve çıkarımda bulunmak çok daha rahat olacaktır.

Bu tez çalışmasında sonuç olarak, otomotiv üretim sektöründe kalite kontrol alanında veri madenciliği yöntemleri ile özgün bir tez çalışması oluşturularak kalite veri seti üzerinde sonucu tahmin etmeye çalışan veri madenciliği modelleri oluşturulmuş ve en iyi performansı sağlayan model bulunmuştur. Ek olarak yine veri madenciliğinin alt alanı olan veri görselleştirme ile karmaşık ve büyük verilere anlam kazandırılarak yorumlanması kolay hale getirilmiştir.

Veri madenciliğinin kalite kontrol sürecinde ne kadar önemli olduğunu anlatarak, bu alanda çalışma yapacaklara yol göstermesi en büyük temennimizdir.

7. KAYNAKÇA

[1] W. Deng, G. Weng "A novel water quality data analysis framework based on time-series data mining," *Journal of Environmental Management*, vol. 196, pp. 365–375, 2017.

[2] A. Baykasoğlu,"Veri madenciliği ve çimento sektöründe bir uygulama", *Akademik Bilişim*, ss. 1-14, 2005.

[3] G. Robert, "A Holistic Approach for quality oriented mainctance planning supported by data mining methods.", *Procedia CIRP*, vol. 57, pp. 259-264, 2016.

[4] J.A. Harding, M. Shahbaz, Srinivas and A. Kusiak "Data Mining in Manufacturing: A Review.", *Journal of Manufacturing Science and Engineering*, vol. 128, no. 4, pp. 969–976, 2005.

[5] A.M.M. Kamal, "A Data Mining Approach for Improving Manufacturing Processes Quality Control", *Next Generation Information Technology*, Gyeongju, South Korea, 2011.

[6] A.R. Khan, H. Schiøler, T. Knudsen "Statistical data mining for efficient quality control in manufacturing", *Emerging Technologies & Factory Automation (ETFA)*, Luxembourg, Luxembourg, 2015.

[7] R. S. Chen, Y.C. Chen and C.C. Chen, "Using data mining technology to deign a quality control system for manufacturing industry", *Advances in Communications, Computers, Systems, Circuits and Devices*, Puerto De La Cruz, Tenerife, 2015.

[8] S. Ferreiro, B. Sierra, I. Irigoien and E. Gorritxateg, "Data mining for quality control: Burr detection in the drilling process.", *Computers & Industrial Engineering*, vol. 60, no. 4, pp. 801-810, 2011.

[9] C. Shearer, "The CRISP-DM model: the new blueprint for data mining.", *Journal of data warehousing*, vol. 5, no. 4, pp. 13-22, 2000.

[10] M. Yılmaz, "Kalite yönetim sistemlerinin evrimi ve toplam kalite yönetiminin banknot matbaası genel müdürlüğünde uygulanabilirliği", Uzmanlık yeterlilik tezi, İşletme Anabilim Dalı, Gazi Üniversitesi, Ankara, Türkiye, 2003.

[11] A.S. Koyuncugil, N. Özgülbaş, "Veri Madenciliği: Tıp ve sağlık hizmetlerinde kullanımı ve uygulamaları", *Bilişim Teknolojileri Dergisi*, c. 2, s. 2, ss. 22-32, 2009.

[12] (Anonim). (2017, 20 Eylül). *Introduction to Data Mining and Knowledge Discovery* [Online]. Erişim: http://www.twocrows.com/intro-dm.pdf.

[13] K. KayaAlp, "Asenkron motorlarda veri madenciliği ile hata tespiti", Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 2007.

[14] M.S. Başarslan, "Telekomünikasyon sektöründe müşteri kayıp analizi", Yüksek lisans tezi, Bilgisayar Mühendisliği Anabilim Dalı, Düzce Üniversitesi, Düzce, Türkiye, 2017.

[15] O. İnan, "Veri Madenciliği", Yüksek lisans tezi, İşletme Anabilim Dalı, Selçuk

Üniversitesi, Konya, Türkiye, 2003.

[16] N.Arıcı, S. Çiftçi, "Uzaktan eğitimde öğrencilerin ders çalışma etkinliklerinin log verilerinin analiz edilerek incelenmesi", *e-Journal of New World Sciences Academy,* vol. 2, no. 4, pp. 1-12, 2006.

[17] H. Turgut, "Yapay zeka uygulamaları", *Ders Notları*, Süleyman Demirel Üniversitesi, Burdur, Türkiye, 2011.

[18] G. Sarıman, "Veri madenciliğinde kümeleme teknikleri üzerine bir çalışma: K-Means ve K-Medoids kümeleme algoritmalarının karşılaştırılması", *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi,* c. 15, s. 3, ss. 8-17, 2011.

[19] Ş. G. Öğüdücü, "Veri madenciliği veri ön işlem", *Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi,* s. 21, ss. 67-76, 2003.

[20] A. Oğuzlar, "Veri Ön İşleme", *Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi,* s. 21, ss. 67-76, 2003.

[21] K. Ergün, "Veri madenciliği veri ön işleme-2", *Ders Notları*, Balıkesir Üniversitesi, Balıkesir, Türkiye, 2009.

[22] EUROMSG. (2016, 9 Ocak). *CRISP Nedir* [Online]. Erişim: http://blog.euromsg.com/crisp-nedir-crisp-surecinde-kullanilan-6-asama/.

[23] A. Gümüşçü, R. Taşaltın, İ. B. Aydilek, "C4.5 karar ağaçlarında genetik algoritma ile budama", *Dicle Üniversitesi Fen Bilimleri Enstitüsü Dergisi,* ss. 77-80, 2016.

[24] E. Uzun. (2014, 3 Eylül). *Naive bayes classifier* [Online]. Erişim: https://www.e-adys.com/makine_ogrenmesi/naive-bayes-classifier/.

[24] E. Ardıl, "Esnek Hesaplama Yaklaşımı İle Yazılım Hata Kestirimi", Yüksek lisans tezi, Bilgisayar Mühendisliği Anabili Dalı,Trakya Üniversitesi, Tekirdağ, Türkiye, 2009.

8. EKLER

8.1. EK 1: C4.5 Çapraz Geçerleme 5 Kat

```
|   |   Deger2 <= 5: 1 (12141.0/1689.0)
|   |   Deger2 > 5
|   |   |   Deger3 <= 4: 1 (550.0/207.0)
|   |   |   Deger3 > 4
|   |   |   |   Durum = GECE
|   |   |   |   |   Personel = P0101: 0 (3.0/1.0)
|   |   |   |   |   Personel = P0488: 0 (494.0/235.0)
|   |   |   |   |   Personel = P0497: 1 (9.0/4.0)
|   |   |   |   |   Personel = P1053: 1 (185.0/86.0)
|   |   |   |   Durum = GÃœNDÃœZ
|   |   |   |   |   Personel = P0101: 1 (0.0)
|   |   |   |   |   Personel = P0488: 1 (244.0/104.0)
|   |   |   |   |   Personel = P0497: 0 (46.0/19.0)
|   |   |   |   |   Personel = P1053: 1 (58.0/28.0)

Number of Leaves  :     12

Size of the tree :      19


Time taken to build model: 0.42 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       14619               85.7621 %
Kappa statistic                          0.6503
Mean absolute error                      0.2231
Root mean squared error                  0.3347
Relative absolute error                 50.0262 %
Root relative squared error             70.8704 %
Total Number of Instances            17046

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,612    0,018    0,945      0,612    0,743      0,680    0,828     0,806     0
                 0,982    0,388    0,833      0,982    0,902      0,680    0,828     0,853     1
Weighted Avg.    0,858    0,264    0,871      0,858    0,848      0,680    0,828     0,837

=== Confusion Matrix ===

     a     b   <-- classified as
  3500  2222 |     a = 0
   205 11119 |     b = 1
```

8.2. EK 1: C4.5 Çapraz Geçerleme 10 Kat

```
|   |   Deger2 <= 5: 1 (12141.0/1689.0)
|   |   Deger2 > 5
|   |   |   Deger3 <= 4: 1 (550.0/207.0)
|   |   |   Deger3 > 4
|   |   |   |   Durum = GECE
|   |   |   |   |   Personel = P0101: 0 (3.0/1.0)
|   |   |   |   |   Personel = P0488: 0 (494.0/235.0)
|   |   |   |   |   Personel = P0497: 1 (9.0/4.0)
|   |   |   |   |   Personel = P1053: 1 (185.0/86.0)
|   |   |   |   Durum = GÃœNDÃœZ
|   |   |   |   |   Personel = P0101: 1 (0.0)
|   |   |   |   |   Personel = P0488: 1 (244.0/104.0)
|   |   |   |   |   Personel = P0497: 0 (46.0/19.0)
|   |   |   |   |   Personel = P1053: 1 (58.0/28.0)

Number of Leaves  :     12

Size of the tree :      19


Time taken to build model: 5.59 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       14628               85.8149 %
Kappa statistic                          0.6506
Mean absolute error                      0.2232
Root mean squared error                  0.3346
Relative absolute error                 50.0526 %
Root relative squared error             70.8462 %
Total Number of Instances            17046

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,607    0,015    0,953      0,607    0,742      0,682    0,827     0,814     0
                 0,985    0,393    0,832      0,985    0,902      0,682    0,827     0,853     1
Weighted Avg.    0,858    0,266    0,873      0,858    0,848      0,682    0,827     0,840

=== Confusion Matrix ===

     a     b   <-- classified as
  3475  2247 |     a = 0
   171 11153 |     b = 1
```

8.3. EK 1: Random Forest Çapraz Geçerleme 5 Kat

```
              Personel
              Tezgah
              Olcu1Adi
              Deger1
              Olcu2Adi
              Deger2
              Olcu3Adi
              Deger3
              Durum
              Sonuc
Test mode:    5-fold cross-validation

=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 10.73 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       14569               85.4687 %
Kappa statistic                          0.6472
Mean absolute error                      0.2024
Root mean squared error                  0.3225
Relative absolute error                 45.3885 %
Root relative squared error             68.3024 %
Total Number of Instances            17046

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,629    0,031    0,910      0,629    0,744      0,669    0,891     0,862     0
                 0,969    0,371    0,838      0,969    0,899      0,669    0,891     0,934     1
Weighted Avg.    0,855    0,257    0,862      0,855    0,847      0,669    0,891     0,910

=== Confusion Matrix ===

     a     b   <-- classified as
  3600  2122 |     a = 0
   355 10969 |     b = 1
```

8.4. EK 1: Random Forest Çapraz Geçerleme 10 Kat

```
              Personel
              Tezgah
              Olcu1Adi
              Deger1
              Olcu2Adi
              Deger2
              Olcu3Adi
              Deger3
              Durum
              Sonuc
Test mode:    5-fold cross-validation

=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 12.05 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       14569               85.4687 %
Kappa statistic                          0.6472
Mean absolute error                      0.2024
Root mean squared error                  0.3225
Relative absolute error                 45.3885 %
Root relative squared error             68.3024 %
Total Number of Instances            17046

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,629    0,031    0,910      0,629    0,744      0,669    0,891     0,862     0
                 0,969    0,371    0,838      0,969    0,899      0,669    0,891     0,934     1
Weighted Avg.    0,855    0,257    0,862      0,855    0,847      0,669    0,891     0,910

=== Confusion Matrix ===

     a     b   <-- classified as
  3600  2122 |     a = 0
   355 10969 |     b = 1
```

8.5. EK 1: Naive Bayes Çapraz Geçerleme 5 Kat

```
  weight sum                5722   11324
  precision                    1       1

Olcu3Adi
  Ã‡AKMA KUVVETÄ°        5723.0 11325.0
  [total]               5723.0 11325.0

Deger3
  mean                   4.033  4.4336
  std. dev.             1.3764  0.7571
  weight sum              5722   11324
  precision                  1       1

Durum
  GECE                  3673.0  7570.0
  GÃœNDÃœZ               2051.0  3756.0
  [total]               5724.0 11326.0



Time taken to build model: 0.8 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       13565               79.5788 %
Kappa statistic                          0.4871
Mean absolute error                      0.3316
Root mean squared error                  0.4018
Relative absolute error                 74.3431 %
Root relative squared error             85.0789 %
Total Number of Instances            17046

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,479    0,044    0,846      0,479    0,612      0,523    0,761     0,722     0
                 0,956    0,521    0,784      0,956    0,861      0,523    0,761     0,817     1
Weighted Avg.    0,796    0,361    0,805      0,796    0,778      0,523    0,761     0,785

=== Confusion Matrix ===

     a     b   <-- classified as
  2740  2982 |     a = 0
   499 10825 |     b = 1
```

8.6. EK 1: Naive Bayes Çapraz Geçerleme 10 Kat

```
  weight sum                5722   11324
  precision                    1       1

Olcu3Adi
  Ã‡AKMA KUVVETÄ°        5723.0 11325.0
  [total]                5723.0 11325.0

Deger3
  mean                     4.033  4.4336
  std. dev.                1.3764  0.7571
  weight sum                5722   11324
  precision                    1       1

Durum
  GECE                     3673.0  7570.0
  GÃœNDÃœZ                2051.0  3756.0
  [total]                5724.0 11326.0



Time taken to build model: 0.08 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       13619               79.8956 %
Kappa statistic                          0.4962
Mean absolute error                      0.3315
Root mean squared error                  0.4017
Relative absolute error                 74.3362 %
Root relative squared error             85.071  %
Total Number of Instances            17046

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,488    0,044    0,849      0,488    0,620      0,531    0,762     0,722     0
                 0,956    0,512    0,787      0,956    0,863      0,531    0,762     0,817     1
Weighted Avg.    0,799    0,355    0,808      0,799    0,782      0,531    0,762     0,785

=== Confusion Matrix ===

     a     b   <-- classified as
  2792  2930 |     a = 0
   497 10827 |     b = 1
```

8.7. EK 1: SMO Çapraz Geçerleme 5 Kat

```
BinarySMO

Machine linear: showing attribute weights, not support vectors.

         0.2808 * (normalized) Personel=P0101
 +      -0.0933 * (normalized) Personel=P0488
 +      -0.0937 * (normalized) Personel=P0497
 +      -0.0937 * (normalized) Personel=P1053
 +      -0.0006 * (normalized) Deger1
 +       1.2491 * (normalized) Deger2
 +       1.5004 * (normalized) Deger3
 +      -0.0007 * (normalized) Durum=GÃœNDÃœZ
 -       0.9057

Number of kernel evaluations: 71054836 (44.778% cached)



Time taken to build model: 50.31 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       11828               69.3887 %
Kappa statistic                          0.1255
Mean absolute error                      0.3061
Root mean squared error                  0.5533
Relative absolute error                 68.6343 %
Root relative squared error            117.1627 %
Total Number of Instances            17046

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,109    0,011    0,839      0,109    0,193      0,228    0,549     0,391     0
                 0,989    0,891    0,687      0,989    0,811      0,228    0,549     0,687     1
Weighted Avg.    0,694    0,595    0,738      0,694    0,604      0,228    0,549     0,588

=== Confusion Matrix ===

     a     b   <-- classified as
   624  5098 |     a = 0
   120 11204 |     b = 1
```

8.8. EK 1: SMO Çapraz Geçerleme 10 Kat

```
BinarySMO

Machine linear: showing attribute weights, not support vectors.

         0.2808 * (normalized) Personel=P0101
 +      -0.0933 * (normalized) Personel=P0488
 +      -0.0937 * (normalized) Personel=P0497
 +      -0.0937 * (normalized) Personel=P1053
 +      -0.0006 * (normalized) Deger1
 +       1.2491 * (normalized) Deger2
 +       1.5004 * (normalized) Deger3
 +      -0.0007 * (normalized) Durum=GÃœNDÃœZ
 -       0.9057

Number of kernel evaluations: 71054836 (44.778% cached)



Time taken to build model: 56.34 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       11828               69.3887 %
Kappa statistic                          0.1255
Mean absolute error                      0.3061
Root mean squared error                  0.5533
Relative absolute error                 68.6343 %
Root relative squared error            117.1627 %
Total Number of Instances            17046

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,109    0,011    0,839      0,109    0,193      0,228    0,549     0,391     0
                 0,989    0,891    0,687      0,989    0,811      0,228    0,549     0,687     1
Weighted Avg.    0,694    0,595    0,738      0,694    0,604      0,228    0,549     0,588

=== Confusion Matrix ===

     a     b   <-- classified as
   624  5098 |     a = 0
   120 11204 |     b = 1
```

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Hikmet CANLI

Doğum Tarihi ve Yeri : 12.01.1995

Yabancı Dili : İngilizce

E-posta : hikmetcnli@hotmail.com

ÖĞRENİM DURUMU

Derece	Alan	Okul/Üniversite	Mezuniyet Yılı
Y. Lisans	Bilgisayar Müh.	Düzce Üniversitesi	2017
Lisans	Bilgisayar Müh.	Düzce Üniversitesi	2015
Lise	Fen Bilimleri	Gölköy Lisesi	2009