



**OTOMOTİV SEKTÖRÜNDE YAKIT TÜKETİMİ VE EMİSYON
SEVİYELERİNİN VERİ MADENCİLİĞİ İLE ANALİZİ**

Simge DENİZ

**YÜKSEK LİSANS TEZİ
ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI**

**GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

ARALIK 2016

Simge DENİZ tarafından hazırlanan “OTOMOTİV SEKTÖRÜNDE YAKIT TÜKETİMİ VE EMİSYON SEVİYELERİNİN VERİ MADENCİLİĞİ İLE ANALİZİ” adlı tez çalışması aşağıdaki jüri tarafından OY BİRLİĞİ ile Gazi Üniversitesi Endüstri Mühendisliği Anabilim Dalında YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Danışman: Prof. Dr. Hadi GÖKÇEN

Endüstri Mühendisliği, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum.

.....

Başkan : Doç. Dr. Mehmet ATAK

Endüstri Mühendisliği, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum.

.....

Üye : Yrd. Doç. Dr. Babek ERDEBİLLİ

Endüstri Mühendisliği, Atılım Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum.

.....

Tez Savunma Tarihi: 08/12/2016

Jüri tarafından kabul edilen bu tezin Yüksek Lisans Tezi olması için gerekli şartları yerine getirdiğini onaylıyorum.

.....

Prof. Dr. Hadi GÖKÇEN
Fen Bilimleri Enstitüsü Müdürü

ETİK BEYAN

Gazi Üniversitesi Fen Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmada;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmada yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Simge DENİZ

08/12/2016

OTOMOTİV SEKTÖRÜNDE YAKIT TÜKETİMİ VE EMİSYON SEVİYELERİNİN VERİ MADENCİLİĞİ İLE ANALİZİ

(Yüksek Lisans Tezi)

Simge DENİZ

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Aralık 2016

ÖZET

Bu tez çalışması otomotiv sektöründe big data uygulama potansiyelini araştırmak amacıyla yazılmıştır. Bu amaca yönelik olarak, yakıt tüketimi ve emisyon seviyesini etkileyen en önemli araç kullanım ve sürüş parametreleri belirlenmiş ve veri madenciliği teknikleriyle sınıflandırılmıştır. Euro 6 emisyon sertifikasyon verisi ve kişisel araçların teknik verileri kullanılmıştır. SPSS programının çeşitli araçları kullanılarak veri seti düzenlenmiş ve analiz edilmiştir. Segmentasyon algoritmaları K-means ve Two-Step kullanılarak SPSS sonuçları analiz edilmiş ve etkili değişkenler belirlenmiş ve sonraki analizlerde kullanılmak üzere kategorize edilmiştir. Her bir parametrenin önemi, yakıt tüketimine katkısını tahmin edilmesi amacıyla veri analitik teknikleri ile değerlendirilmiştir. Böylece, yakıt verimliliği için optimal kontrol stratejileri geliştirilmesi mümkün olacaktır. Bu çalışmada öngörü kabiliyeti ve uygulama kolaylığı açısından sınıflandırma tekniklerinden yapay sinir ağları (neural network), bayes ağları ve C5.0 algoritması tercih edilmiştir. Bu teknikler Clementine 12.0 programında uygulandığında genel olarak araç ağırlığının ve motor kapasitesinin yakıt tüketimi, emisyon seviyesi üzerinde en etkili parametreler olduğu bulunmuştur. Yöntemler Clementine'daki Evaluation Node ile değerlendirilmiş ve elde edilen grafikler incelendiğinde en iyi tahmini C5.0 algoritmasının yaptığı bulunmuştur. Fakat, bazı durumlarda yapay sinir ağları daha iyi ve güvenilir sonuçlar vermektedir.

Bilim Kodu : 90602

Anahtar Kelimeler : Veri Analitiği, Büyük Veri, Yakıt Ekonomisi, Emisyon, Segmentasyon, Performans Değerleme

Sayfa Adedi : 118

Danışman : Prof. Dr. Hadi GÖKÇEN

APPLICATION OF DATA MINING IN AUTOMOTIVE INDUSTRY FOR
ANALYZING FUEL CONSUMPTION AND EMISSION LEVELS

(M. Sc. Thesis)

Simge DENİZ

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

December 2016

ABSTRACT

This thesis is aimed to investigate application potential of big data mining in the automotive industry. The most important usage and driving parameters, which effects fuel consumption and emission level of passenger cars identified and classified by using data mining methods. A dataset created by combining Euro 6 and technical specification data of passenger cars has been analyzed using different tools of SPSS, such as, descriptive statistics, correlations, regression and etc. The results have been compared and effecting parameters have been derived by segmentation algorithms aiming better results by categorizing variables for upcoming analysis. The importance of each parameter has been evaluated to predict its contribution on fuel consumption by data analytics technics. Therefore, it will be possible to build optimal control strategies for fuel efficiency. Adopted data mining technics in this study are classification algorithms, such as, neural networks, Bayesian networks and C5.0 algorithm as well as segmentation algorithms (e.g. K-means and Two-step) targeting foreseeability and simplicity. Application of those technics by Clementine 12.0 has shown that weight and engine capacity of passenger cars were the most important parameters in fuel consumption, respectively. Depending on the evaluation of the performance of those methods by Evaluation Node of Clementine, it has been found that C5.0 was the most efficient method to predict fuel consumption among others. However, the evaluation charts (Gain, Profit, ROI, etc.) have shown that neural network could have better prediction results in some conditions.

Science Code : 90602

Key Words : Data Analytics, Big Data, Fuel Efficiency, Emissions, Segmentation, Performance Evaluation

Page Number : 118

Supervisor : Prof. Dr. Hadi GÖKÇEN

TEŞEKKÜR

Başta akademik bilgi ve tecrübelerini paylaşarak yol gösterdikleri için danışmanım Sayın Prof.Dr. Hadi Gökçen'e ve Erasmus öğrenim hareketliliğim süresince Karlsruhe Teknoloji Enstitüsü'nde danışmanlığımı üstlenen Sayın Prof.Dr. Gholamreza Nakhaeizadeh'ye olmak üzere, Karlsruhe Teknoloji Enstitüsü'ne gidişime vesile olan Sayın Doç.Dr. Mehmet Atak'a, bu tezin hazırlanmasında maddi ve manevi desteğini esirgemeyen Peiman Alipour Sarvari'ye, tezin biçimsel düzenlemesinde zaman kazandıran Sayın Kudret Ekici, Farhad Hanifepour ve Esra Özkan'a ve her zaman çocukları için elinden gelenin en iyisini yapmaya gayret eden sevgili aileme teşekkürü bir borç bilirim.

Ayrıca; ümitsizliğe düştüğüm her an bana hedeflerimi hatırlatan Dr. Canan Dağdeviren ile zamanından çok önce gelen filozof Friedrich Wilhelm Nietzsche'ye ve sahip olduğum birçok hakkı borçlu olduğum Mustafa Kemal Atatürk'e saygı ile teşekkürlerimi sunarım.

İÇİNDEKİLER

	Sayfa
ÖZET	iv
ABSTRACT.....	v
TEŞEKKÜR.....	vi
İÇİNDEKİLER	vii
ÇİZELGELERİN LİSTESİ.....	x
ŞEKİLLERİN LİSTESİ.....	xi
HARİTALARIN LİSTESİ.....	xv
SİMGELER VE KISALTMALAR.....	xvi
1. GİRİŞ.....	1
2. OTOMOTİV SEKTÖRÜNDE BÜYÜK VERİ	9
2.1. Büyük Veri	9
2.1.1. Büyük veride 5V	9
2.2. Otomotiv Sektörünün Büyük Veriden Etkilendiği Alanlar	11
3. LİTERATÜR ARAŞTIRMASI	15
4. METODOLOJİ.....	19
4.1. Veri Madenciliği.....	19
4.1.1. Veri madenciliği prosedürü (CRISP DM)	19
4.2. Makine Öğrenmesi	21
4.2.1. Kümeleme (segmentasyon)	22
4.2.2. Sınıflandırma	23
4.3. Çalışmada Kullanılan Metodoloji.....	29
5. UYGULAMA	31
5.1. Veri.....	31

	Sayfa
5.1.1. Veri üretimi	31
5.1.2. Betimleyici istatistikler	31
5.2. Tasarım Aşaması	40
5.3. Analiz Aşaması.....	40
5.3.1. Şehir içi yakıt tüketimi	41
5.3.2. Şehir dışı yakıt tüketimi	43
5.3.3. Kombine yakıt tüketimi	45
5.3.4. Karbon monoksit (CO) salınımı.....	46
5.3.5. Karbon dioksit (CO ₂) salınımı	48
5.3.6. Hidrokarbon (THC) emisyonu	49
5.3.7. Nitrojen oksit (NO _x) emisyonu.....	51
5.3.8. Parçacık emisyonu.....	53
5.3.9. Gürültü seviyesi.....	54
5.4. Algoritma Performanslarının Değerlendirilme Aşaması	58
5.4.1. Şehir içi yakıt tüketimi	59
5.4.2. Şehir dışı yakıt tüketimi	64
5.4.3. Kombine yakıt tüketimi.....	67
5.4.4. Karbon monoksit (CO) salınımı	70
5.4.5. Karbon dioksit (CO ₂) salınımı.....	73
5.4.6. Hidrokarbon (THC) emisyonu	76
5.4.7. Nitrojen oksit (NO _x) emisyonu.....	79
5.4.8. Parçacık emisyonu.....	83
5.4.9. Gürültü seviyesi.....	86

	Sayfa
6. SONUÇLAR	91
6.1. Faktörlerin Değerlendirilmesi	91
6.2. Benchmarking	93
6.3. Tartışma.....	97
6.4. Sonuçların Değerlendirilmesi.....	98
KAYNAKLAR	99
EKLER.....	103
EK-1. Üreticilerin dağılımını gösteren histogram	104
EK-2. Şanzıman tiplerinin dağılımı	105
EK-3. Her bir üreticinin motor kapasitesini gösteren kutu diyagramları.....	106
EK-4. Test çevrimlerine göre yakıt tüketimi dağılımları	107
EK-5. Şehir dışı yakıt tüketimi (a) yükseltme (b) ROI grafiği	108
EK-6. Kombine yakıt tüketimi için (a) yükseltme (b) ROI grafiği.	109
EK-7. Karbon monoksit emisyonu için (a) yükseltme (b) ROI grafiği.	110
EK-8. Karbon dioksit emisyonu için (a) Yükseltme (b) ROI grafiği.....	111
EK-9. Hidrokarbon emisyonu için (a) Yükseltme (b) ROI grafiği	112
EK-10. Parçacık emisyonu için (a) Yükseltme (b) ROI grafiği.....	113
EK-11. Gürültü seviyesi için (a) Yükseltme (b) ROI grafiği.....	114
EK-12. Şehir içi yakıt tüketimi için (a) yükseltme (b) ROI grafiği (Benchmarking).....	115
ÖZGEÇMİŞ	116

ÇİZELGELERİN LİSTESİ

Çizelge	Sayfa
Çizelge 2.1. Test ve müşteri araç filolarından senaryolarla tahmin edilen veri kümelerinin boyutları.....	9
Çizelge 5.1. Üreticilerin dağılımı.....	33
Çizelge 5.2. Motor kapasitesinin istatistikleri.....	36
Çizelge 5.3. Yakıt tiplerinin dağılımı.....	38
Çizelge 6.1. Sınıflandırma tekniklerinin tahmin başarısı.....	91
Çizelge 6.2. Sinir ağı ile bulunan parametre ağırlıkları	92
Çizelge 6.3. C5.0 algoritması ile bulunan parametre ağırlıkları	92
Çizelge 6.4. Bayes ağı ile bulunan parametre ağırlıkları	93
Çizelge 6.5. Özetlenmiş benchmarking sonuçları.....	96

ŞEKİLLERİN LİSTESİ

Şekil	Sayfa
Şekil 1.1. 2035 yılı için simülasyon sonuçları	4
Şekil 1.2. Her üreticinin ortalama araç ağırlığına bağlı CO ₂ emisyon hedefleri	4
Şekil 1.3. T-Systems şirketinin Big Data sistemi	6
Şekil 2.1. Büyük veride 5V (Leboeuf, 2016).....	9
Şekil 4.1. CRISP DM referans modelinin altı aşaması.....	20
Şekil 4.2. Bazı makine öğrenmesi teknikleri	21
Şekil 4.3. Gizli katmanlı bir sinir ağı.....	26
Şekil 4.4. Ağırlıklandırılmış sinir ağı (x düğümünden y düğümüne kadar olan ağırlık)	26
Şekil 4.5. Sinir ağında bulunan tek bir birimdeki işlemler	27
Şekil 4.6. Uygulanan metodolojinin işlem adımları	29
Şekil 5.1. Üreticilerin dağılımı.....	32
Şekil 5.2. Şanzıman tipleri	35
Şekil 5.3. Motor kapasitesi dağılımı	37
Şekil 5.4. Her test aracına kurulan portatif emisyon ölçüm sistemi (PEMS)	41
Şekil 5.5. Şehir içi yakıt tüketimi için (a) iki aşamalı kümeleme algoritması ve (b) k-ortalamlar algoritması ile belirlenen kümeler.....	42
Şekil 5.6. Şehir içi (a) iki aşamalı kümeleme algoritması ve (b) k-ortalamlar algoritması ile bulunan kümelerin dağılımlarının karşılaştırılması.....	42
Şekil 5.7. Şehir dışı yakıt tüketimi için (a) iki aşamalı kümeleme algoritması ve (b) k-ortalamlar algoritması ile belirlenen kümeler.....	44
Şekil 5.8. Şehir dışı yakıt tüketimi için (a) iki aşamalı kümeleme algoritması ve (b) k-ortalamlar algoritması ile bulunan kümelerin dağılımlarının karşılaştırılması	44
Şekil 5.9. Kombine yakıt tüketimi için (a) iki aşamalı kümeleme algoritması (b) k-ortalamlar algoritması ile belirlenen kümeler.....	45

Şekil	Sayfa
Şekil 5.10. Kombine yakıt tüketimi için (a) iki aşamalı kümeleme algoritması ve (b) k-ortalamlar algoritması ile bulunan kümelerin dağılımlarının karşılaştırılması	46
Şekil 5.11. Karbon monoksit (CO) salınımı için (a) iki aşamalı kümeleme algoritması (b) k-ortalamlar algoritması ile belirlenen kümeler	47
Şekil 5.12. Karbon monoksit (CO) salınımı için (a) iki aşamalı kümeleme algoritması ve (b) k-ortalamlar algoritması ile bulunan kümelerin dağılımlarının karşılaştırılması	47
Şekil 5.13. Karbon dioksit (CO ₂) salınımı için (a) iki aşamalı kümeleme algoritması (b) k-ortalamlar algoritması ile belirlenen kümeler	48
Şekil 5.14. Karbon dioksit (CO ₂) salınımı için (a) iki aşamalı kümeleme algoritması ve (b) k-ortalamlar algoritması ile bulunan kümelerin dağılımlarının karşılaştırılması	49
Şekil 5.15. Hidrokarbon (THC) emisyonu için (a) iki aşamalı kümeleme algoritması (b) k-ortalamlar algoritması ile belirlenen kümeler	50
Şekil 5.16. Hidrokarbon (THC) emisyonu için (a) iki aşamalı kümeleme algoritması ve (b) k-ortalamlar algoritması ile bulunan kümelerin dağılımlarının karşılaştırılması	50
Şekil 5.17. Standartlardaki emisyon limitleri ve yolda ölçülen değerler arasındaki fark	51
Şekil 5.18. Nitrojen Oksit (NO _x) emisyonu için (a) iki aşamalı kümeleme algoritması (b) k-ortalamlar algoritması ile belirlenen kümeler	52
Şekil 5.19. Nitrojen oksit (NO _x) emisyonu için (a) iki aşamalı kümeleme algoritması ve (b) k-ortalamlar algoritması ile bulunan kümelerin dağılımlarının karşılaştırılması	52
Şekil 5.20. Parçacık emisyonu için (a) iki aşamalı kümeleme algoritması (b) k-ortalamlar algoritması ile belirlenen kümeler	53
Şekil 5.21. Parçacık emisyonu için (a) iki aşamalı kümeleme algoritması ve (b) k-ortalamlar algoritması ile bulunan kümelerin dağılımlarının karşılaştırılması	54
Şekil 5.22. M1 tip özel araçlarının ses seviyesi kategorileri.....	55
Şekil 5.23. Gürültü seviyesi için (a) iki aşamalı kümeleme algoritması ve (b) k-ortalamlar algoritması ile belirlenen kümeler	56

Şekil	Sayfa
Şekil 5.24. Gürültü seviyesi için (a) iki aşamalı kümeleme algoritması ve (b) k-ortalamlar algoritması ile bulunan kümelerin dağılımlarının karşılaştırılması	56
Şekil 5.25. Şehir içi yakıt tüketimi için (a) kazanç (b) kâr grafiği.....	60
Şekil 5.26. Nitrojen oksit (NO _x) emisyonu için (a) iki aşamalı kümeleme algoritması ve (b) k-ortalamlar algoritması ile bulunan kümelerin dağılımlarının karşılaştırılması	61
Şekil 5.27. Nitrojen oksit (NO _x) emisyonu için (a) iki aşamalı kümeleme algoritması ve (b) k-ortalamlar algoritması ile bulunan kümelerin dağılımlarının karşılaştırılması	63
Şekil 5.28. Nitrojen oksit (NO _x) emisyonu için (a) iki aşamalı kümeleme algoritması ve (b) k-ortalamlar algoritması ile bulunan kümelerin dağılımlarının karşılaştırılması	63
Şekil 5.29. Nitrojen oksit (NO _x) emisyonu için (a) iki aşamalı kümeleme algoritması ve (b) k-ortalamlar algoritması ile bulunan kümelerin dağılımlarının karşılaştırılması	64
Şekil 5.30. Şehir dışı yakıt tüketimi için tepki grafiği	65
Şekil 5.31. Şehir dışı yakıt tüketimi verilerine uygulanan tekniklerin analiz sonuçları .	66
Şekil 5.32. Şehir dışı yakıt tüketimini etkileyen (a) sinir ağı (b) Bayes ağı (c) C5.0 algoritması ile bulunan en önemli faktörler	66
Şekil 5.33. Kombine yakıt tüketimi için (a)kazanç (b) kâr grafiği	67
Şekil 5.34. Kombine yakıt tüketimi için tepki grafiği	68
Şekil 5.35. Kombine yakıt tüketimi verilerine uygulanan tekniklerin analiz sonuçları..	69
Şekil 5.36. Kombine yakıt tüketimini etkileyen (a) sinir ağı (b) Bayes ağı (c) C5.0 algoritması ile bulunan en önemli faktörler	70
Şekil 5.37. Karbon monoksit salınımı için (a) kazanç (b) kâr (c) yanıt grafiği	71
Şekil 5.38. Karbon monoksit (CO) verilerine uygulanan tekniklerin analiz sonuçları...	72
Şekil 5.39. Karbon monoksit (CO) salınımını etkileyen (a) sinir ağı (b) Bayes ağı (c) C5.0 algoritması ile bulunan en önemli faktörler	73
Şekil 5.40. Karbondioksit (CO ₂) salınımı için (a)kazanç (b) kâr (c) yanıt grafiği.....	74

Şekil	Sayfa
Şekil 5.41. Karbon dioksit (CO) salınımı verilerine uygulanan tekniklerin analiz sonuçları	75
Şekil 5.42. Karbon dioksit (CO ₂) salınımını etkileyen (a) sinir ağı (b) Bayes ağı (c) C5.0 algoritması ile bulunan en önemli faktörler	76
Şekil 5.43. Hidrokarbon (THC) salınımı için (a) kazanç (b) kâr (c) yanıt grafiği	77
Şekil 5.44. Hidrokarbon (THC) emisyon verilerine uygulanan tekniklerin analiz sonuçları	78
Şekil 5.45. Hidrokarbon (THC) emisyonunu etkileyen (a) sinir ağı (b) Bayes ağı (c) C5.0 algoritması ile bulunan en önemli faktörler	79
Şekil 5.46. Nitrojen oksit (NO _x) salınımı için (a) kazanç (b) kâr grafiği	80
Şekil 5.47. Nitrojen oksit (NO _x) emisyonu için (a) tepki, (b) yükseltme ve (c) ROI grafikleri	81
Şekil 5.48. Nitrojen oksit (NO _x) emisyon verilerine uygulanan tekniklerin analiz sonuçları	82
Şekil 5.49. Nitrojen Oksit (NO _x) emisyonunu etkileyen (a) sinir ağı (b) Bayes ağı (c) C5.0 algoritması ile bulunan en önemli faktörler	82
Şekil 5.50. Parçacık emisyonu için (a) kazanç (b) kâr (c) yanıt grafiği	83
Şekil 5.51. Parçacık emisyonu verilerine uygulanan tekniklerin analiz sonuçları	85
Şekil 5.52. Parçacık emisyonunu etkileyen (a) sinir ağı (b) Bayes ağı (c) C5.0 algoritması ile bulunan en önemli faktörler	85
Şekil 5.53. Gürültü seviyesi için (a) kazanç (b) kâr (c) yanıt grafiği	86
Şekil 5.54. Gürültü seviyesi verilerine uygulanan tekniklerin analiz sonuçları	88
Şekil 5.55. Gürültü seviyesini etkileyen (a) sinir ağı (b) Bayes ağı (c) C5.0 algoritması ile bulunan en önemli faktörler	88
Şekil 6.1. Tekniklerin karşılaştırma sonuçları	94
Şekil 6.2. Şehir içi yakıt tüketimi için (a)kazanç (b) kâr (c) yanıt grafiği	95

HARİTALARIN LİSTESİ

Harita	Sayfa
Harita 1.1. 2030'da hafif araç stoğu	3



SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler

cm³	santimetre küp
db	desibel
km	kilometre
l	litre
mpg	mili galon
s	saniye

Kısaltmalar

BN	Bayes ağı
CO	Karbon monoksit
CO₂	Karbon dioksit
CRISP DM	Veri madenciliği prosedürü (Cross-Industry Standardization Procedure-Data Mining)
IoT	Şeylerin İnterneti (Internet of Things)
NN	Sinir ağı
THC	Toplam hidrokarbon

1. GİRİŞ

Gelecek yıllarda yeni trendler otomotiv sektöründeki üreticiler, bayilikler, müşteri ve sürücüler üzerinde etki yaratacaktır. KPMG şirketi tarafından 2014'te yayınlanan Global Automotive Executive Survey'e göre sektör daha global bir hal alacak ve araç üreticileri esnek ve modüler platformlar kullanmak zorunda kalacaktır (King, 2007).

Pazarların küreselleşmesi, kaynaklar ve bilginin artması ile ürün geliştiren şirketlerinin müşteri ihtiyaçlarına ve çevresel değişimlerdeki sorumluluğu artmaktadır. Otomotiv sektöründe, karbondioksit emisyonlarının sınırlandırılmasındaki güçlüklerle karşın, genişleyen daha heterojen pazarlarda daha fazla müşteriye yüksek kalitede ürünler sunulması beklenmektedir. Bunun için müşteri ihtiyaçlarını belirlemede ve ürün geliştirmeye katkıda bulunan performans ve istatistik değerlerinin elde edilmesinde çok verimli araçlara ve metotlara ihtiyaç duyulmaktadır. Telematik hizmetler kullanılan araç filolarından toplanan verilerin hacimce büyük oluşu nedeniyle, sistemleri ölçülebilen ve verimli bir şekilde tasarlamak oldukça önemlidir. Bunu otomotivde Big Data ile tarif etmek mümkündür (Johanson, Belenki, Jalminger, Fant, ve Gjertz, 2015).

Bilgi ve iletişim teknolojilerindeki gelişmelerden dolayı segmentasyon tekniklerinin seçimi önem kazanmaktadır. Özellikle de veritabanı yönetim sistemleri ve veri madenciliği pazarlamanın yönünü değiştirmektedir. Yeterli ölçüde verinin bulunabilirliği ve geleneksel istatistik tekniklerinin (veya istatistik tabanlı segmentasyon araçlarının) böyle büyük veriler üzerinde etkin olmayan performansı sebebiyle, araştırmacılar pazarlar ve müşterileri hakkında kullanışlı bilgi elde etmek için daha etkili segmentasyon araçları kullanmaya yönelmiştir. Böylece, bilgi keşfi ve veri madenciliği daha iyi bir çözüm olarak görülmüştür (Hiziroglu, 2013).

Harvard Business Review ("Does Your Company Have a Data Science Strategy?", 2016)'e göre, uzmanların işletme problemlerini ve veri analitiklerinin etkili bir strateji yürütebilmesi için hangi soruları yanıtlayabileceğini açıkça tanımlaması gerekmektedir. Bunun yapılmaması, verilerin şirketi yanlış yöne sürüklemesine neden olacak ve bu şirketler için risk yaratacaktır. Algoritmaları aşama aşama veriden öğrenen bir çeşit yapay zekâ çeşidi olan makine öğrenmesi, verileri nerede araması gerektiği programlanmadan dahi veri hakkında fikir elde edebilmektedir. Bu, büyük miktarlardaki veriler üzerinde çalışmayı

kolaylaştırmaktadır. Makine öğrenmesinin sunduğu bilgi edinme hızıyla şirketler sahip oldukları verileri stratejik olarak kullanmada ve uzmanların daha bilinçli kararlar vermesinde yardımcı olacaktır.

Otomobil, ulaşım ihtiyaçlarının giderilmesi için petrol ve diğer yakıtları kullanmaktadır. Yakıta olan bu bağılıktan dolayı çevre ve enerji ile ilgili problemlerle doğrudan bağlantılıdır. Çünkü toplam petrol kullanımının 2/3'ü ve ülkedeki karbon emisyonunun 1/3'ü ulaşım sektöründen kaynaklanmaktadır. Bu yüzden, enerji güvenliği ve global iklim değişikliği ile ilgili kaygılarla, sonraki yıllarda yeni otomobillerin yakıt tasarrufunun geliştirilmesi için yasal öneriler sunulmuştur (Lynette, Christopher, Anup, ve John, 2007). Akıllı ve verimli ulaşım sistemleri göz önüne alındığında araç sistemlerinde yakıt verimliliği önem kazanmakta ve otomotivde Big Data'nın kaynak olarak kullanılabileceği bir potansiyel geliştirme alanı olarak ortaya çıkmaktadır. Günümüzde, tanınmış otomotiv şirketleri otonom(sürücüsüz) araçlar, elektrikli araçlar gibi daha ekonomik, daha çevreci, teknolojik ve kişiselleştirilmiş otomobiller üretmeye yönelmişlerdir. Bu araçların kullanım parametrelerinin yakıt tüketimi ve emisyon üzerindeki etkisi incelendiğinde ve anlaşıldığında, üreticilerin belirledikleri hedeflere ulaşması ve sürücülere daha ekonomik sürüş imkanı sağlamak mümkün olacaktır.

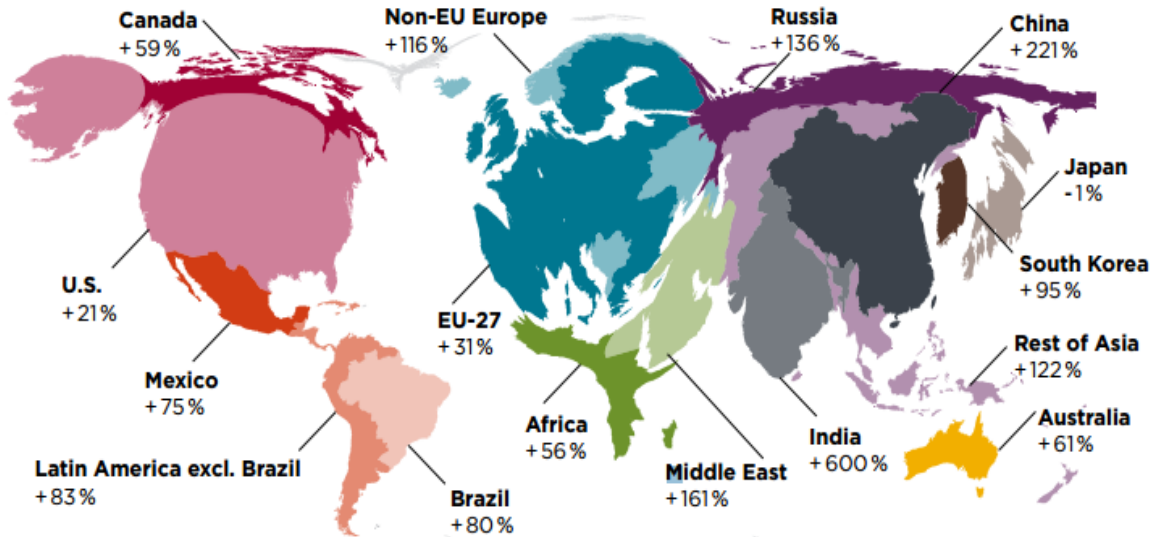
Bu tezde, otomobillerin Euro 6 ve teknik özellik bilgilerinden oluşturulan bir veri seti hazırlanarak yakıt tüketimi ve emisyon seviyeleri üzerinde etkili parametreler analiz edilmiştir. Yakıt verimliliğini arttırmak, emisyon ve ses seviyelerini düşürmek amacıyla otomotiv sektöründe büyük veri madenciliğinin uygulanma potansiyeli araştırılmıştır. Böylece, yakıt verimliliğinde kullanılabilecek optimal kontrol stratejileri geliştirmek mümkün olacaktır.

Bu çalışmada, büyük miktarlardaki veriye uygulanmaya elverişli olan veri analitiği ve makine öğrenmesi teknikleri incelenmiştir. Çalışmanın sonuçları kullanılarak veri analitiği araçları çeşitli senaryolarla karşılaştırılabilir. Sonuçlar araçların çevresel sertifikasyonu açısından tasarım aşamasını geliştirmeye yardımcı olacaktır. Aynı zamanda, bu sonuçların matematiksel modellerde kullanılması araç rotalama problemlerini ve simülasyon bazlı değerlendirmeleri geliştirecektir.

Altı bölümden oluşan bu tezin ilk kısmında konu tanıtılmakta, 2.kısmında çalışmanın amaçlarından, otomotiv sektörü ile ilgili bilgiler ve emisyon ve yakıt verimliliğine yönelik yeni trendlerden bahsedilmektedir. Önceki çalışmalarda ele alınan benzer problemlerde kullanılan veri madenciliği yaklaşımları ve big data uygulamalarına değinilmektedir. 3.bölümde, CRISP DM prosedürü, Big Data Analitiği ve veri madenciliği tekniklerinden söz edilmekte ve 4.bölümde tezde kullanılan veri özellikleri açıklanmaktadır. Tekniklerin bu veri üzerinde uygulanma adımları 5.bölümde verilmiştir ve sonuçlar yorumlarıyla beraber 6.bölümde özetlenmektedir.

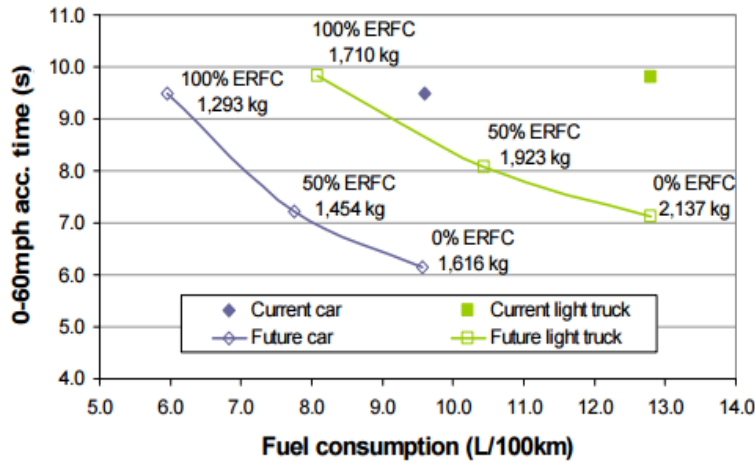
Otomotiv sektörü ve yeni trendler

2010 yılı verilerine göre dünya genelinde yaklaşık olarak 1 milyar otomobil ve kamyon bulunmaktaydı. Avrupa Birliği Araç Pazar İstatistikleri'ne göre (ICCT Europe, 2013), 2030 yılı için bu rakamın 1.7 milyara ulaşması ve en çok büyüyecek olan pazarların Asya ve Orta Doğu olacağı beklenmektedir. Bu büyüme ve 2010 yılı verileri ile karşılaştırılması Harita 1.1'de gösterilmiştir.



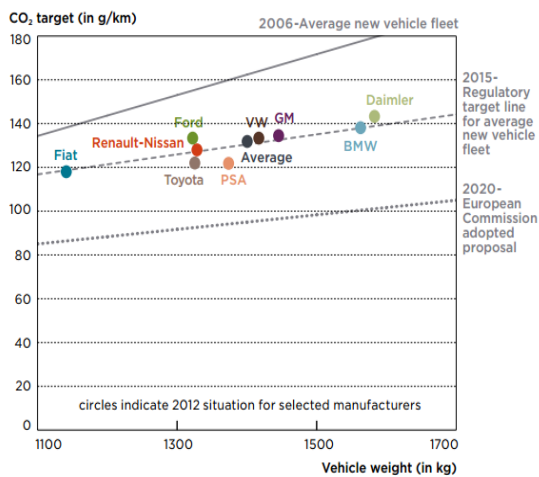
Harita 1.1. 2030'da hafif araç stoğu (ICCT Europe, 2013)

Bir simülasyon yazılımı ile yapılan analizler göstermiştir ki, gelişen araç teknolojilerine bağlı olarak geleceğin araçlarının yakıt tüketimi seviyesi yükselirken hızlanma süresi azalacaktır (Şekil 1.2).



Şekil 1.1. 2035 yılı için simülasyon sonuçları (Lynette ve diğerleri, 2007)

2012’de Avrupa Komisyonu tarafından hedeflenen ortalama CO₂ salınımı 2020 için 95 g/km, bunun yakıt tüketim karşılığı da 3.8 litre/100 km olarak belirlenmiştir. 2025 için belirlenen CO₂ hedefi ise 68-78 g/km olmuştur (ICCT Europe, 2013). Avrupa Birliği yaptırımları kapsamında, her üretici için CO₂ emisyon hedefleri belirli araçların ağırlığına göre uyarlanmıştır (Şekil 1.2). Bu yüzden 2015 ile 2020 yılları arasında gerekli olan CO₂ salınımını (ve dolayısı ile yakıt kullanımını) düşürme oranı her üretici için eşitlenmiş ve %27 olarak belirlenmiştir. 2014’te sınırları %68 (mazot karbon monoksit) ile %96 (dizel parçacıklar) arasında değişen Euro 6 standardı etkili olmaya başlamıştır.



	2012	2015	2020
Daimler	143 g/km 1583 kg	140 g/km	101 g/km
BMW	138 g/km 1563 kg	139 g/km	100 g/km
GM	134 g/km 1445 kg	133 g/km	96 g/km
Volkswagen	133 g/km 1417 kg	132 g/km	96 g/km
All manufacturers (average)	132 g/km 1400 kg	131 g/km	95 g/km
Ford	129 g/km 1322 kg	128 g/km	92 g/km
Renault-Nissan	128 g/km 1329 kg	128 g/km	93 g/km
Fiat (incl. Chrysler)	124 g/km 1209 kg	123 g/km	89 g/km
Toyota	122 g/km 1325 kg	128 g/km	92 g/km
PSA (Peugeot-Citroën)	122 g/km 1374 kg	130 g/km	94 g/km
Fiat (excl. Chrysler)	118 g/km 1141 kg	119 g/km	86 g/km

Şekil 1.2. Her üreticinin ortalama araç ağırlığına bağlı CO₂ emisyon hedefleri (ICCT Europe, 2013)

Bu hedeflere ulaşmak için, Lynette ve diğerleri (2007) hâlihazırda bulunan yakıt tasarrufu teknolojilerini değerlendirmiş ve ilgili üretim maliyetlerini çıkarmışlardır. Sera gazı emisyonu ve karbondioksit üzerindeki etkilerini yaşam döngüsü bazında değerlendirmiş ve otomobilde yakıt tüketiminin azaltılması için çeşitli senaryolar canlandırmışlardır. Yakıt verimliliğini arttırmak için kullandıkları yaklaşımlar konvansiyonel araç teknolojisi, alternatif powertrains ve araç ağırlığının azaltılması olmuştur.

Big Data analizleri de yakıt maliyetlerini düşürmede kullanılabilecek en iyi imkânlardan biridir ve şirketlerin rekabette avantaj sağlamasında fırsatlar sunmaktadır. Sürekli veriler, bunların doğru analizi ve temiz dizel enerjisi ile birlikte ele alınması maliyet tasarrufu yaratacak ve karbon ayak izini azaltan bir katma değer sağlayacaktır (Fleet Owner, 2009). Otomobiller birbirine ve internete bağlandığında diğer otomobil ve yol kullanıcıların hareketlerini izleyebilecek ve gerektiğinde müdahale edebilecektir. Bir trafik kazasının meydana geleceğini önceden bilen araçlar, farklı bir rota ve bu rota üzerinde yakıt tüketimine yönelik optimizasyon bile sunabilecektir (Rijmenam, n.d.). Örneğin; Mercedes me connect için sağlanan Canlı Trafik Bilgisi benzer trafik kazalarını önlemeye yardımcı olmakta varılacak istikamete mümkün olan en kısa sürede ulaşmaya yardımcı olmaktadır. Sürücü hareketleri gelişmiş bir trafik veri sisteminde gerçek zamanlı olarak kaydedilmektedir. Bu da trafik bilgisini araçlara göndermeyi mümkün kılmaktadır (Mercedes-Benz, 2016).

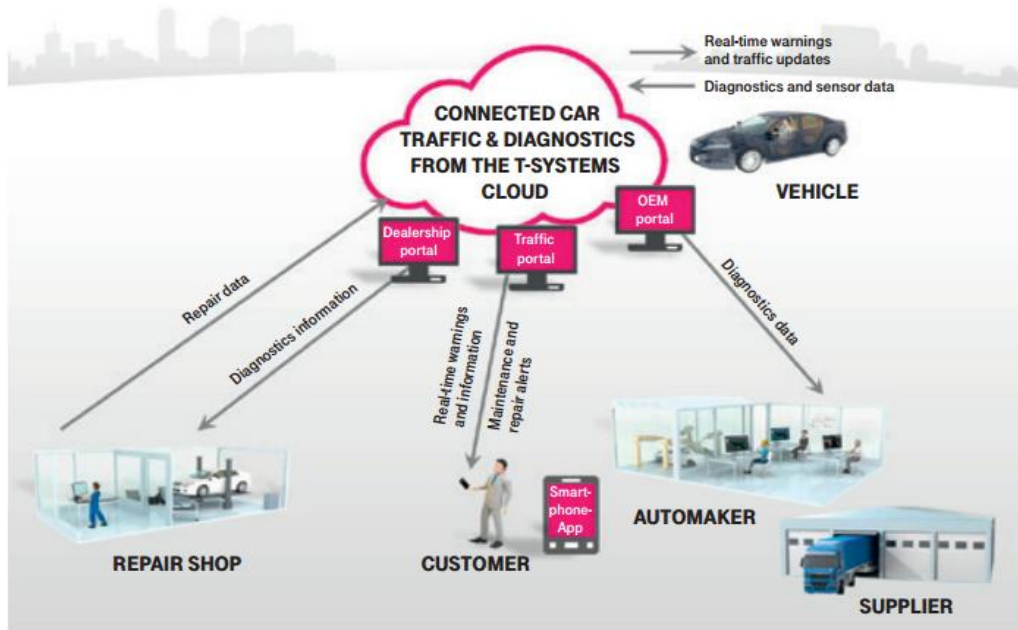
Burada temel düşünce, bu aktivitelerin connectivity ve veri analitikleri ile kontrol edilerek yakıt kullanımının etkin bir biçimde azaltılabilesidir (Anand, 2014). Yakıt verimliliği, araç verimliliğinin ve sürücü davranışlarının geliştirilmesi ile arttırılabilir. Bu da veri toplanması ve analiz edilmesi ile, geliştirmelerde kullanılması uygun parametrelerin belirlenmesiyle ve araç sahibinin sürüş anında SMS veya araçtaki göstergeler ile bilgilendirilmesi ile yapılabilir.

Araçları internete bağlayarak bu işlemleri gerçekleştirebilen Automatic, Mojio ve Zubie gibi cihazlar mevcuttur (Fowler, 2015). Bunlar motor, yakıt verimliliği, aracın konumu gibi verileri toplamaktadır. Apple CarPlay ve Android Auto'dan farklı olarak bu cihazlar, araç içinde neler olduğuna dair bilgileri de ortaya çıkarmaktadır. Uzun geliştirme periyotları, önceki teknolojilere bağlılık, arızalara karşı tedbir gereklilikleri ve teknoloji şirketlerinin hizmet ilişkilerini değiştirme korkusu sebebiyle otomobil üreticileri şimdiye kadar akıllı telefon tipi bağlanma şekline yavaş ilerlemiştir. 2015 Lincoln ve Chevrolet modelleri de

dahil olmak üzere bazı yeni araçlar kablosuz bağlantılarla gelmeye başlamaktadır. Ford gibi otomobil üreticileri dış uygulamalar için araba verilerini kullanıma açmaktadır.

Towers (2016)'a göre, Ford Otomotiv Şirketi daha iyi iş modelleri ve bunun çevresindeki potansiyel yeni işler geliştirmek için yollardaki milyonlarca otomobilden veri elde etmektedir. Planları sadece iş modellerini geliştirmekle kalmayıp, araçlarının çevresindeki yapıyı (yollar, şehirler ve diğer akıllı otomobiller) da geliştirmek hedeflenmektedir. Bu işleme başlamak için Ford araç içi sensörlerle 200 çalışanının sürüş alışkanlıkları üzerinde çalışmaktadır. Bunun altında yatan düşünce araç içinde neler olduğunu ve nasıl sürüldüğünü izlemek ve hangi metriklerin ve izlemelerin en etkili olduğuna karar vermektir.

Ford'un karşılaşması muhtemel zorluk ise buldukları birçok görüşün gerçekte uygulanmasıdır. Bu kadar verinin Ford'a filtrelenerek tekrar iletilmesi, çok yüksek seviyelerde veri işleme gücü ve istedikleri etkiyi yaratacak örüntü ve trendleri bulma yetenekleri gerektirecektir.



Şekil 1.3. T-Systems şirketinin Big Data sistemi (T-Systems, 2013)

T-Systems International GmbH (2013) otomobil şirketlerine Big Data çözümleri sunmaktadır. Connected Car ve Diagnostics aracıyla, otomobildeki bir adaptör sürücünün cep telefonunda yüklenmiş bir uygulama ile sensör değerlerini iletmektedir (Şekil 1.3).

Kablosuz bağlantıyla bu değerlerin online olarak işlendiği ve dağıtıldığı T-Systems bulut sistemine gönderilmektedir. Böylece araç üreticileri hem ürün geliştirmede hem de maliyetli ürün geri çağırımlarının önlenmesinde kullanılmak üzere bilgi zenginliğine ulaşmaktadır.

XL Hybrid şirketi de benzer şekilde XL Link bulut tabanlı büyük veri analitiği sistemini duyurmuştur. Bu ürün yakıt performansını da içeren anahtar performans göstergelerini ölçmekte ve müşterilerin hibrit-elektrikli filo araçlarında düşürülen karbon dioksit salınımlarını bildirmektedir (Fleet Owner, 2016).

Bilgi teknolojileri sunucusu Hewlett-Packard (HP) ve otomotiv tedarikçisi IAV (Ingenieurgesellschaft Auto und Verkehr), Connected otomobilin akıllı ulaşım konseptlerinin nasıl optimal şekilde kullanılabileceğine dair fikir edinmek için test sürüşlerini kaydederek büyük veri analizi yapmışlardır. Toplanan veriler HP Helion Bulut sisteminde analiz edilmiştir (Kielman ve Dettling, 2015).

IBM Analytics (IBM Corp., 2016) de büyük veri uygulamalarında otomotiv şirketleri için bir çözümdür. Örneğin; Honda IBM'in Internet of Things for Automotive çözümünü kullanarak güç birimlerinden ürettiği veriyi gerçek zamanlıya yakın analizler için doğrudan bulut sistemine göndermektedir. Honda, Formula 1 araçlarının güç birimlerindeki 160'tan fazla sensörden aldığı veriyi yönetmek ve analiz etmek için 2016'da 'Şeylerin İnterneti' teknolojisini uygulamaya başlamaktadır (O'Connor, 2016).

Audi büyük veriye dayalı pilot projeler üzerinde çalışmaktadır (AutomotiveIT, 2013). İş zekâsı yapıları ve uyumluluğu zaten önceden duyurulmuştu. Audi'nin amacı, Ford ile benzer şekilde var olan çekirdek işi güçlendirmek ve büyük veriyi kullanarak yeni iş modelleri geliştirmektir.

Volvo Grup Kamyonları yakıt verimliliği konusuna odaklanmış ve farklı perspektiflerden bakılmasına gerek duyulduğu fark edilmiştir. Sistem yapısının ve alternatif yakıtların müşteri ve hizmet sağlayıcıları gibi diğer paydaşlarla işbirliği yapılması gerektiği görülmüştür (Johnsen ve Felldin, 2016). Bunun sonucunda, Volvo araçlarının kullanım ve performans bilgilerini içeren kodlanmış araç verisi (Logged Vehicle Data-LVD) kullanılarak yakıt tüketiminde etkisi olabilecek parametreler kümesi analiz edilmiştir.

Çalışmanın amacı

Bu çalışmada, veri madenciliğinin büyük veri üzerinde uygulama potansiyeli araştırılmıştır. Büyük hacimlerdeki veriye uygulanması amaçlanan veri analizi ve makine öğrenmesi yöntemleri, üç farklı sürüş tarzının koşullarında modellenmiş ve özel araçların gerçek hayattaki yakıt tüketimi üzerinde etkili değişkenlerin önemini bulmada kullanılmıştır. Bu çalışmanın sonuçları kullanılarak, çeşitli veri analitiği araçlarının performanslarının farklı senaryolara bağlı olarak karşılaştırılması mümkün olacaktır.

Bu çalışmanın sonuçları, çevresel sertifikasyon açısından araçların tasarım aşamasının geliştirilmesine yardımcı olacaktır. Diğer yandan, bu tezde elde edilen sonuçların matematiksel modellerde kullanılması araç rotalama problemlerinin ve simülasyon bazlı değerlendirmelerin daha iyi şekilde geliştirilmesini sağlayacaktır.

2. OTOMOTİV SEKTÖRÜNDE BÜYÜK VERİ

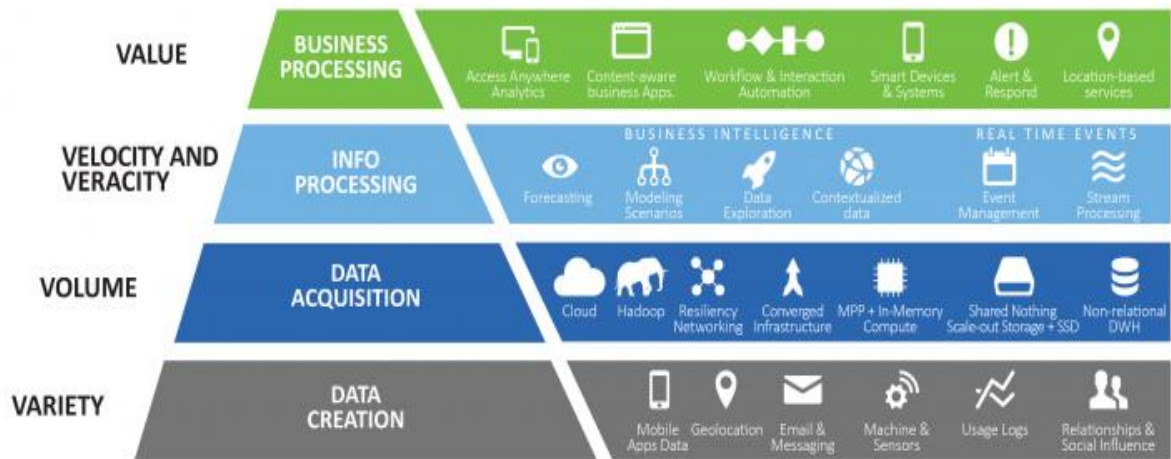
2.1. Büyük Veri

Büyük Veri (Big Data) yeni yeni popüler olmuş, veri tabanı yönetimi sistemlerinde ve geleneksel veri işleme araçlarında kullanılırken başa çıkılması zor, çok büyük ve kompleks veri kümelerini temsil eden bir terimdir. Otomotiv alanında ise, araçların bağlantılı oluşu (connectivity) ve telematik hizmetlerin kullanıldığı araçlardan geniş çapta veri yakalanması ve bu verinin analizi olarak yer almaktadır (Johanson ve diğerleri, 2015).

2.1.1. Büyük veride 5V

2001’de önceki adı MetaGroup olan Gartner analist şirketi tarafından üç boyutlu veriler için 3V; hacim (volume), hız (velocity) ve çeşitlilik (variety) olarak tanımlanmıştır. Sonradan iki yeni boyut, değer (value) ve bulunurluğu (veracity) eklenmiştir.

Büyük Veriyi anlayabilmek için geliştirilen 5V modelinin boyutları Johanson ve diğerleri (2015) tarafından otomotiv veri uygulama senaryolarında Şekil 2.1’deki gibi açıklanmıştır.



Şekil 2.1. Büyük veride 5V (Leboeuf, 2016)

Otomotiv Verisinin Hacmi (Volume): Johanson ve diğerleri (2015) tarafından geliştirilen çeşitli senaryolarla tahmin edilen veri hacmi Çizelge 2.1’deki gibi bulunmuştur.

Çizelge 2.1. Test ve müşteri araç filolarından senaryolarla tahmin edilen veri kümelerinin boyutları

Uygulama	Filo	Günlük	Aylık
CAN otobüs işletimi	Müşteri	560 TB	206 PB
CAN otobüs işletimi	Test	4.5 TB	1.6 PB
Remote Diagnostic Read-Out	Müşteri	100 GB	36 TB
Remote Diagnostic Read-Out	Test	10 GB	3.6 TB
Mevcut Sağlık Durumu	Müşteri	1 GB	365 GB
Mevcut Sağlık Durumu	Test	100 MB	36 GB

Google’ın sürücüsüz araçları şimdiden gerçek bir veri üreticisidir. Otomobilin sürücü olmadan sürülmesini sağlayan tüm sensörlerle her saniye yaklaşık olarak 1 gigabayt veri üretilmektedir (Rijmenam, n.d.). Amerikalılar otomobillerini yılda 600 saat kullanmaktadır. Bunun her bir araç için yılda 2.160.000 saniye ve yaklaşık 2 petabayt veriye eşit olduğu ve araç sayısının dünya çapında bir milyona ulaşacağı düşünülürse, üretilen verinin miktarı tahmin edilemeyecek boyutlara ulaşmaktadır.

Otomotiv üreticisi Audi günde gigabaytlık boyutlarda veriyle çalışmaktadır (AutomotiveIT, 2013). Ancak bu alandaki veri miktarı birçok faktöre bağlıdır ve çeşitlenebilir. Müşterilerin kullanım davranışı ve araçtaki hizmet portfolyosu (pazar ve modele bağlı olarak) kararı etkilemektedir. Bundan dolayı bu alanda doğru bir şekilde ölçümlemek önemlidir. Audi’nin Audi Connect Center’da yaptığı gibi özel bulut teknolojileri ile bu amaca ulaşılabilir.

Otomotiv Verisinin Hızı (Velocity): Veri yüksek bir hızda akmaktadır. Büyük veri zamana karşı duyarlı olmasından dolayı kimi zaman bir dakika dahi çok geç olabilmektedir. Birçok kurumun veri hızı en büyük zorluktur. Kredi kartı transferleri ve sosyal medya mesajları mili saniyede yapılmakta ve buradan üretilen veriler veri tabanlarına kaydedilmektedir (Kumar ve Ravishankar, 2015).

Otomotiv Verisinin Çeşitliliği (Variety): Aynı alanda veya farklı yerde olsa bile, veri kaynakları inanılmaz derecede heterojendir. Her tipte dosya çeşitli formatlarda gelmekte ve yazı, ses, log dosyaları, videolar vb. şekilde yapılandırılmış ya da yapılandırılmamış olabilir. Çeşitliliğin sınırı yoktur ve veri, bağlantıya sayılmadan ve nitelendirilmeden girer. Hadoop

gibi sistemler sadece hacim ve çeşitlilik gibi kısımların üstesinden gelebilir fakat gerçek zamanlı büyük veri analitiği tamamını kapsar (Kumar ve Ravishankar, 2015).

Otomotiv Verisinin Bulunurluğu (Veracity): Bulunurluk, müşteri araçlarında ve bağlantılı hizmetlerde güvenliği garantilemek için otomotiv alanı ile yüksek ilişkiye sahiptir. Bilgi bazlı ürün geliştirmede faydalanılması için yakalanan verinin alakası ve kalitesinden emin olmak için provenance ve izlenebilirlik çok önemlidir.

Bulunurluğun bu alandaki diğer bir yönü de otomotiv şirketlerinin mühendislik verileri konusunda gizliliğe eğilim göstermeleridir. Çünkü bu veriler, halka açık ağ yapılarında iletişim sağlarken karmaşık bilgi güvenliği mekanizmaları gerektirmektedir (Johanson ve diğerleri, 2015).

Otomotiv Verisinin Değeri (Value): Büyük Otomotiv Verisi hizmetleriyle mümkün kılınan çeşitli faydalar mevcuttur. Yeni satış sonrası hizmetler ve ürün eklentileri, bulut tabanlı veri işleme ve yönetim servisleri kullanılarak bağlantılı araçlardan alınan, bütünleştirilen ve analiz edilen bilgi kaynaklarına bağlı bir biçimde tasarlanabilir. Bu hizmetler, koruyucu ve önleyici bakım hizmetlerini, aktif güvenlik ve otonom sürüş destek hizmetlerini vb. kapsamaktadır (Johanson ve diğerleri, 2015).

2.2. Otomotiv Sektörünün Büyük Veriden Etkilendiği Alanlar

Büyük veri, konsept tasarımı, belgelendirme, üretim, pazarlama, finans, performans geliştirme, hizmet geliştirme ve satış sonrası vb. çeşitli amaçlarla kullanılabilir. Büyük verinin otomotiv sektörü üzerindeki etkileri detaylı bir biçimde alt başlıklar halinde özetlenmiştir (CSC, 2016).

Konsept Tasarımı: Otomobil üreticileri, aracın kullanım ömrü boyunca toplanan verilerden elde edilen Büyük Veriyi gelecekteki araçların tasarımını şekillendirmede kullanacaktır. Milyonlarca millik sürüşlerden alınan gerçek dünya verileri, güvenlik, aerodinamik, performans, güç algoritmaları ve aracın diğer temel özelliklerini etkileyecektir.

Belgelendirme: Büyük Veri ile optimize edilen tedarik zinciri yönetimi, belgelendirme sürecinde yeni verimlilik ölçütlerine ulaşılmasını sağlayacaktır.

Üretim: Montaj hattında üretim süreci boyunca elde edilen veriler, imalat simülasyonlarını geliştirmek ve sonraki montaj hatlarını daha verimli ve esnek yapan makine performansını sağlamak için öngörücü analitiklerde kullanılacaktır.

Pazarlama: Büyük Veri şimdiden araç pazarlamasındaki en büyük etkiye sahiptir. Sosyal çevre, üreticilerin yeni araçlar tasarlama planlarında artan bir rol oynayacaktır. Mevcut modeller ile ilgili müşteri geri bildirimleri de pazarlama uzmanlarına anahtar temalar ve yeni kampanyalar için mesajlar belirlemesinde yardımcı olacaktır.

Finans: Müşteri alışkanlıklarının, tercihlerinin ve Pazar segmentlerindeki alış gücü, üreticilere daha etkili finans programları geliştirmesinde fikir vermektedir. Bu yalnızca ilk adım olmakla birlikte, satış ve sahadaki kullanım verileri üzerinde yapılan Büyük Veri Analizleri finans şirketlerinin yeni hizmetler ve yeni kazanç akışları geliştirmelerine yardımcı olacaktır.

Performans: Gerçek hayattaki araç performansı hem Büyük Veriyi etkileyecek hem de ondan faydalanacaktır. Sahadan (araç sistemleri, sürücü giriş verileri, dış şartlar vb.) toplanan bilgi, parçaların tasarımında ve gelecekteki araçların karakteristikleri üzerindeki en büyük etkiyi yaratacaktır.

Hizmet: Otomotiv sektöründe, performans gibi hizmet alanında da Büyük Veriye hem katkıda bulunulacak hem de faydalanılacaktır. Milyonlarca hizmet etkinliği boyunca toplanan bilgiler tasarımcılara değerli bir geri bildirim sağlayacaktır. Diğer yandan, hizmet süreci basitleştirilecek ve sürücüler yoldaki kazalardan korunacaktır. Öngörücü analizler sayesinde araçtaki problemler, parçalar ve sistemlerdeki işlev bozuklukları ortaya çıkmadan belirlenebilecektir.

Satış Sonrası: Büyük Veri ile konum bazlı özelliklerin kombinasyonundan kazanılan fikirlere bağlı olarak, sürücülere yapılan satış sonrası hizmetlerin patlaması beklenmektedir. Veri işlenecek ve özel teklif ve eklentileri isteyen sürücülere ulaşacak işbirlikleri geliştirilecektir. Seyahat hizmetleri, sigorta, petrol, bakım ve eğlence bu endüstrilerden yalnızca birkaçıdır.

Günümüzde Büyük Veri kullanan otomotiv şirketleri müşteri ihtiyaçlarına ve kalite konularına odaklanmaktadır. Örneğin; Porsche Büyük Veriyi müşteri taleplerini karşılamak ve onlara daha iyi bir satış sonrası hizmet sunabilmek için kullanmaktadır (MHP, 2014). Araçların verilerini toplamak ve analiz etmek ayrıca önleyici bakım, Radyasyon Terapili Kalite Güvence (Radiation Therapy Quality Assurance) ve tedarik zincirlerinde iş zekası alanlarında kaliteyi arttırmaya yardımcı olmaktadır. Gelecekte, saha verileri toplanarak üreticilere sürücü profilleri çıkarılacak ve müşterilere mobilite ve bağlantılılık hizmetleri sunulacaktır. Diğer bir odak noktası ise Endüstri 4.0'ı destekleyerek üretimi geliştirme olacaktır. Ancak, henüz iş dünyasında Büyük Verinin başarılı uygulamaları olmamıştır (MHP, 2014).

Diğer yandan, günümüzde Büyük Veri potansiyelinin kullanılmasında veri karmaşıklığı, hesaplama karmaşıklığı ve sistem karmaşıklığı açısından birçok zorlukla karşılaşmaktadır. En düşük seviyede işleyen sistemlerin tasarımından, en yüksek seviyede bilimsel araştırmalardaki bir seri açık probleme ve analizlere kadar bu zorluklar etkisini göstermektedir (Jin, Wah, Cheng ve Wang, 2015).

Modern otomobillerin ve yeni sistemlerin geliştirmesinde bazı güvenlik kaygıları ortaya çıkmıştır ve bununla ilgili olarak Büyük Veriyi kullanan e-Call Acil Servis Sistemi (e-Call Emergency System) geliştirilmiştir (Brüninglinghaus, 2014). Ayrıca trafik kanununda yeni maddelere ihtiyaç vardır.



3. LİTERATÜR ARAŞTIRMASI

Önceki çalışmalarda, yakıt tüketimi ve emisyon seviyelerini etkileyen faktörler madencilik sektöründe araştırılmıştır (Dindarloo ve Siami-Irdemoosa, 2016; Wang ve diğerleri 2015). Büyük veri ile bir analiz de taşıma sektöründe yapılmıştır (Fujitsu Laboratories Ltd, 2016). Ahmed (2014) büyük veri araçlarını ve tekniklerini iki veri seti (saatlik enerji kullanım verisi ve cihaz seviyesi verisi) kullanarak binalardaki enerji verimliliği konusunda çalışmıştır. Bu çalışmada clustering ve sınıflandırma teknikleri de kullanılmıştır.

Dindarloo ve Siami-Irdemoosa (2016) çalışmasında seçilen bir madencilik operasyonundaki ağır maden kamyonlarının her bir işlem döngüsündeki yakıt tüketimini tahmin etmek istenmiştir. Çalışma sonuçları kullanıldığında alternatif senaryolar için yakıt maliyetlerinin hesaplanabileceği, böylece yakıt tüketimi ve sera gazı emisyonunu azaltan bir senaryonun seçilebileceğini belirtmişlerdir.

Hafif özel araçların yakıt tüketiminin modellenmesi, gerçek dünyaya uyarlanma şartlarındaki belirsizliklerden kaynaklanan kaygılar yaratmıştır. Bununla ilgili olarak; Zhang ve diğerleri (2016) gerçek hayattaki yakıt tüketimini ve karbon dioksit salınımını tahmin etmek için ampirik bir model geliştirmişlerdir. Yerel filo konfigürasyonu ve işlem şartları dikkate alınarak Macao'daki mazotlu özel araçlarda kullanılmıştır.

Zhang ve diğerleri (2015) yakıt tüketimi ve karbon emisyonunu dikkate alarak bir araç rotalama problemi üzerinde çalışmıştır. Rota ayırma metoduna bağlı olarak model çözümünde geliştirilmiş bir tabu search algoritması kullanmışlardır.

Johanson ve diğerleri (2015) bilgiye dayanan ürün geliştirme için Büyük Otomotiv Verisinin işlenmesindeki fırsatları araştırmışlardır. Connected araçlardan elde edilen verileri yakalamak ve online analizini gerçekleştirmek için bir teknolojik çerçeve sunmuşlardır.

Sürüş modlarında yakıt tüketimi ile emisyon arasında yakın bir ilişki olduğunu iddaa eden diğer bir çalışmada (Li ve Sun, 2014), tek ve çoklu araçlarda numerik simülasyonlar gerçekleştirilmiştir. Burada trafik akış oranı, aracın ortalama hızı, yakıt tüketimi, karbon monoksit (CO), hidro karbon (HC) ve nitrojen bileşikleri (NOx) emisyonları araştırılmıştır.

Farklı metodolojilerle elde edilen yakıt tüketimi ve çevre kirliliğine sebep olan emisyon öngörülleri tartışılmıştır. Aynı şekilde uygun değerlendirme yapılabilmesi için yerel sürüş şartlarının karakterize edilmesinin önemi de araştırılmıştır (Zamboni ve diğerleri, 2015). Özellikle, araç sürüş şartlarından çok fazla etkilenen nitrojen bileşikleri için standart sistemlerin davranışı dikkate alınmıştır.

On-board sürüş destek sistemleri, sürücülerin davranışını ve anlık yakıt tüketimlerini analiz etmekte ve bu davranışlarını ekonomik bir biçimde düzenlemelerine yardımcı olan bir geri bildirim sağlanmaktadır. Zeng, Miwa, ve Morikawa (2016) tarafından yürütülen bir çalışmada taktik seviyesinde benzer bir eko-rotalama stratejisi çalışılmıştır.

Fetene ve diğerleri (2015) gerçek hayattaki sürüş örüntülerini dikkate alarak tamamen akülü elektrikli araçların enerji tüketim oranını etkileyen faktörleri araştırmıştır. Burada dikkate alınan gerçek hayat koşulları; değişik hava koşulları, sürücü karakteristikleri ve yol karakteristikleri olup 741 sürücünden alınan altı aylık sürüş örüntü verisini oluşturmaktadır.

CGI Grubu (CGI GROUP INC., 2014), yakıt tüketimi ile sürüş davranışı arasındaki ilişkiyi araştırmak için Scania'nın zengin sürücü verilerini analiz etmiştir. Bu analizde yapay sinir ağları, CHAID ve lineer regresyon modelleri gibi istatistiksel modeller ve clustering teknikleri kullanılmıştır. Bu analizin amacı ölçülen sürüş davranışını anlık yakıt tüketimine bağlayan bir model geliştirmektir.

Araç sertifikasyon verileri kullanılarak, Duarte ve diğerleri (2015) tarafından sürüş esnasında gerekli olan güce bağlı araç enerji karakterizasyonunu sağlayan bir model geliştirilmiştir. Bu metodoloji geçerlenmiş ve Portekiz'deki çok satan araçların farklı periyotlardaki enerji etkilerini ölçmede kullanılmıştır. Bu sayede bireysel halk verileri kullanılarak yakıt kullanımı ve kirli emisyon profillerini çıkarabilen araç teknolojisinden bağımsız bir yöntem geliştirilmiştir.

Siarni-Irdemoosa ve Dindarloo (2015)'ya göre, her bir döngüdeki yakıt tüketiminin tahmini ve modellenmesi enerji maliyetlerinin değerlendirilmesinde çok önemli bir araçtır ve çoklu regresyonun uygulanmasında bazı önemli istatistiksel varsayımlar mevcuttur. Bu çalışmada da yapay sinir ağları kullanılarak her operasyon döngüsündeki yakıt tüketimi tahmin edilmiş ve bazı önemli parametreler bulunmuştur.

Martin ve diğerleri (2014) Birleşik Krallık'taki özel araçların yakıt tüketimini ölçen Cambridge Otomotiv Araştırma Modelleme Uygulaması (the Cambridge Automotive Research Modelling Application -CARma) adlı yeni bir metodoloji sunmuştur. Bu yöntemde özel araçların araç kütlesi, motor büyüklüğü ve sıkıştırma oranı vb. temel tasarım girdileri kullanılmıştır.

Zhang ve diğerleri (2014) de Çin'deki hafif özel araçların gerçek hayattaki yakıt tüketimi ve karbondioksit emisyon faktörlerini değerlendirmiştir. Etkili trafik ölçülerine bakarak yoldaki sürüş koşulları ve yakıt tüketiminin azaltılmasının önemini tartışmışlardır. Bu çalışmanın sonuçlarıyla, kanun koyucuların gerçek hayattaki trafik şartlarının yakıt tüketimi ve karbondioksit emisyonu üzerindeki rolünü anlamasına yardımcı olunacağı iddaa edilmiştir. Aynı amaçla Ehsani, Ahmadi ve Fadaei (2016) sıcaklık, sürüş şekli, asfalt verimliliği ve yakıt verimliliği gibi yeni parametrelerin etki derecesini araştırmıştır. Yerçekiminin azaldığı, hızlanma anında, dönüşlere dayanıklılık, aerodinamik dayanıklılık ve köşe kayıpları şeklinde beş parçadan oluşan bir model önermişlerdir.

Wang, Zhang ve Ouyang (2015) ise üç çeşit elektrikli aracın sürüş örüntülerini inceleyerek ve bunları konvansiyonel araçlarla karşılaştırarak yakıt kullanımındaki azalmayı değerlendirmiştir. Düşük hızların, uzun kayıp zamanların, ani hız değişimlerinin ve kısa sürüş aralıklarının elektrikli araçları konvansiyonel araçlara göre yakıt tüketimi açısından daha avantajlı yapan sürüş örüntüleri olduğunu bulmuşlardır. Ayrıca, 20-50 kilometrelik mesafelerde kullanılabilecek küçük akülerin sürücü beklentilerini sağlayacağını ve aracın maliyetini düşürmeye yardımcı olacağını bulmuşlardır.

Li, Wang ve Song (2016) araç takip sürecindeki yakıt tüketimini akıllı elektrikli araçlar için optimize etmeye çalışmışlardır. Bir optimizasyon algoritması önermiş ve MATLAB'da simülasyon yaparak sonuçları değerlendirmişlerdir. Önerilen yöntemle yakıt tüketimini azaltmada başarı elde etmişlerdir.

Constantinescu, Marinoiu ve Vladiu (2010) trafik güvenliğini arttırmak için filo yönetimi verilerini kullanarak sürüş örüntülerini araştırmıştır. Geliştirdikleri GPS-tabanlı Gipix adlı bir cihazla şehiriçi trafiğinden sürüş parametrelerini belirlemiş, verileri sunuculardan topladıktan sonra hiyerarşik cluster analizi gibi yöntemler kullanarak analiz etmişlerdir.

Sürüş ortamının, trafik yapısının, sürücünün o anki koşullarının ve karakteristiklerinin sürüş davranışında rol oynadığı bulunmuştur.

Wang ve diğerleri (2015) Bayezyen Olasılık ve Kernel Yoğunluk Tahmini yöntemlerini kullanarak bir istatistiksel örüntü tanıma tekniği geliştirmiştir. Bu çalışmada, sürüş şekillerinin belirsizliği dikkate alınmış ve sürücüler kategorize edilmiştir. Yeni modelin faydaları da çapraz geçirme tekniği ile gösterilmiştir. Önerilen tekniğin bulanık mantık algoritmasından daha sabit sonuçlar verdiği gözlemlenmiştir.

Bu çalışmaların dışında, her bir aracın belirlenen bir yoldaki gerçek zamanlı yakıt kullanım verileri birçok şirket tarafından toplanmakta ve çeşitli emisyon modelleri kullanılarak analiz edilmektedir. Böylece yakıt tüketiminin ve karbon dioksit emisyonunun tahmininde bu modellerin ne kadar etkili olduğu araştırılmaktadır (Bkz. OpenXC, Mercedes me connect vb.).

4. METODOLOJİ

4.1. Veri Madenciliği

Veri madenciliği, büyük hacimli veri yığınları içerisinde karar alabilmek için potansiyel olarak faydalı olabilecek, uygulanabilir ve anlamlı bilgilerin çıkarılmasına verilen addır. Veri madenciliği geniş anlamda veri analiz teknikleri bütünüdür.

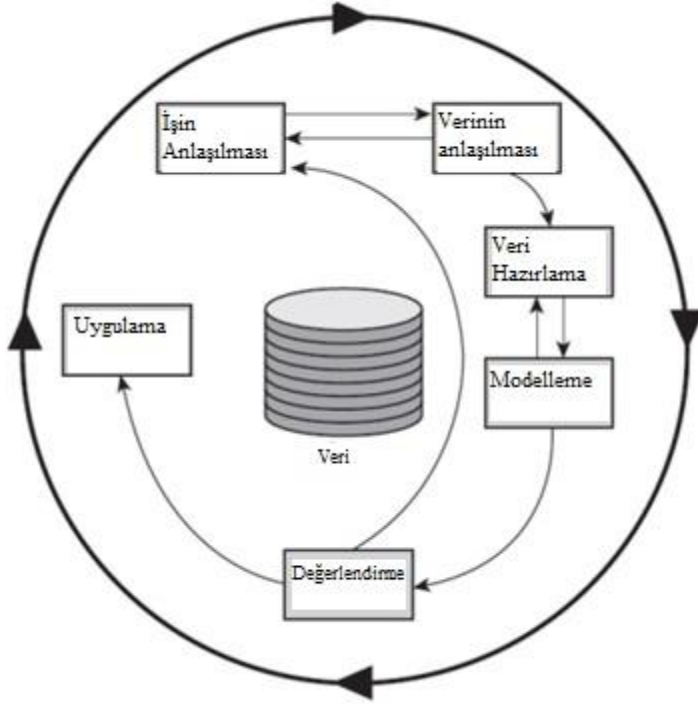
Mevcut problemleri çözmek, kritik kararları almak veya geleceğe yönelik tahminleri yapmak için gerekli olan bilgileri elde etmeye yarayan bir araçtır. Ortaya çıkarılması hedeflenen bilgiler; üstü kapalı, çok net olmayan, önceden bilinmeyen anlamlı ve kritik bilgilerdir (Argüden ve Erşahin, 2008).

4.1.1. Veri madenciliği prosedürü (CRISP DM)

CRISP DM (Cross-Industry Standard Process for Data Mining) veri madenciliği işlemlerine rehberlik eden endüstri onaylı bir yöntemdir (IBM Corp., 2011).

Metodoloji olarak, bir projenin tipik aşamalarının tanımlamalarını, her aşamadaki görevleri ve bu görevler arasındaki ilişkilerin açıklamalarını içerir.

Süreç modeli olarak CRISP DM veri madenciliği yaşam döngüsünün bir genel bakışını sağlar. Yaşam döngüsü modeli, aşamalar arasındaki en önemli bağımlılıkları içeren okların olduğu altı aşamayı içermektedir (Şekil 4.1) Aşamaların sırası değişebilmektedir. Bazı projeler geriye gidebilmekte veya önemli aşamalar arasında tekrarlanabilmektedir.



Şekil 4.1. CRISP DM referans modelinin altı aşaması (Chapman ve diğerleri, 2000)

Bu altı aşama, aşağıdaki gibi veri madenciliğinin daha geniş iş alıştırmalarına katmayı da içeren tüm veri madenciliği sürecini kapsar.

İşin Anlaşılması: Veri madenciliğinin en önemli aşaması sayılabilir. İşletme amaçlarının tanımlanması, durumun değerlendirilmesi, veri madenciliği hedeflerinin konulması ve bir proje planının çıkarılmasından oluşur.

Verinin Anlaşılması: Veri, veri madenciliğinin başlangıç malzemelerini sağlar. Bu aşamada veri kaynaklarının ne olduğuna ve karakteristiklerinin anlaşılmasına olan ihtiyaca işaret edilmektedir. Bu aşamada, başlangıç verisi toplanır, tanımlanır, araştırılır ve veri kalitesi doğrulanır.

Veri Hazırlama: Veri kaynakları tanımlandıktan sonra madencilikte kullanılmak üzere hazırlanır. Gerekli veriler seçilir, temizlenir, yapılandırılır, entegre edilir ve formatlanır.

Modelleme: Bu aşama, veriden bilgi çıkarılmasında analiz yöntemlerinin kullanıldığı, veri madenciliğin can alıcı parçasıdır. Bu aşamada modelleme tekniklerinin seçimi, test tasarımları ve modellerin kurulup değerlendirilmesi yapılır. Değerleme için, çok kriterli karar

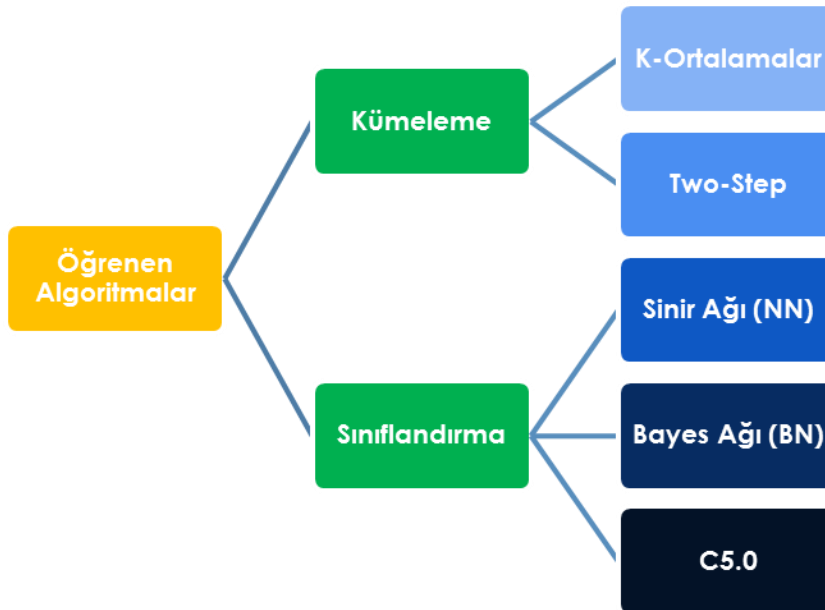
verme ve veri zarflama analizlerine dayalı bazı yaklaşımlar geliştirilmiştir (Nakhaeizadeh ve Schnabl, 1997).

Değerlendirme: Model seçildikten sonra veri madenciliği sonuçlarının işletme amaçlarına nasıl yardımcı olabildiği değerlendirilebilir. Bu aşamada sonuçlar değerlendirilir, veri madenciliği süreci gözden geçirilir ve sonraki aşamalara karar verilir.

Uygulama: Bu aşama işletme probleminin çözülmesi için yeni bilginin günlük işletme süreçlerine entegre edilmesine odaklanır. Plan uygulanması, yönetim ve bakım, final raporunun üretilmesi ve projenin gözden geçirilmesini kapsar (SPSS, 2007).

4.2. Makine Öğrenmesi

Makine öğrenmesi, örnek veriler veya geçmiş tecrübeler kullanarak bir performans kriterini optimize etmek için bilgisayarları programlamaktır. Bazı parametrelere göre tanımlanmış bir model vardır ve öğrenme, eğitim verilerini veya geçmiş deneyimi kullanarak model parametrelerini optimize etmek için bir bilgisayar programının yürütülmesidir. Model, tahminlerde bulunmak için öngöründe bulunabilir veya veriden bilgi edinmek için açıklayıcı olarak kullanılabilir veya her ikisini birden yapabilir (Alpaydın, 2010).



Şekil 4.2. Bazı makine öğrenmesi teknikleri

Makine öğrenmesi, matematiksel modeller oluşturmak için istatistik teorisini kullanır, çünkü temel görev bir örnekten çıkarım yapmaktır. Bazı makine öğrenmesi teknikleri kümeleme ve sınıflandırma başlıkları altında Şekil 4.2'deki gibi incelenmiştir.

4.2.1. Kümeleme (Segmentasyon)

Kümeleme bir veri tabanını farklı gruplara böler (Edelstein, 1999). Kümelemenin amacı birbirinden farklı ve elemanları çok benzer olan gruplar bulmaktır. Gruplanan örnekleri sık sık çıkararak kümelemede değişiklik yapmak gerekmektedir. Çünkü kullanıcı bunları ilgisiz veya anlamsız görmektedir. Veri tabanını mantıklı bir biçimde segmente eden kümeleri bulduktan sonra, bu kümeler yeni verilerde kullanılabilir.

Bölümleme (partition) algoritmaları ve kümelemenin adımları Mooi, Erik, Sarstedt (2011) tarafından açıklanmıştır.

Bu çalışmada, K-ortalamlar ve iki aşamalı (Two-Step) algoritmaları kullanılmıştır.

İki Aşamalı Kümeleme Algoritması (Two-Step Algorithm)

Algoritmanın ismi, yaklaşımın iki basamakta gerçekleşmesinden gelmektedir: İlk aşamada k-ortalamlar algoritmasına benzer bir prosedür takip edilmektedir. Bunun sonuçlarına bağlı olarak iki aşamalı algoritma, homojen kümeler elde etmek için nesneleri sırasıyla birleştirir ve yaprakları yabancı nesneleri gösteren bir küme ağacı çıkartılır (Mooi, Erik; Sarstedt, 2011). Bu prosedür kategorik ve sürekli değişkenlerle çalışabilir ve kullanıcıya küme numaraları ve maksimum küme sayısını belirlemede veya küme sayısını otomatik olarak istatistiksel değerlendirme kriterlerine göre seçmesinde esneklik sağlar. Dahası, belirli bir kümenin oluşturulması için her değişkenin önemini belirler. Bu özellikler iki aşamalı kümeleme algoritmasını geleneksel yöntemlere göre üstün gelmektedir.

k-Ortalamlar Algoritması

Hata kareleri kriterini işleme alan en basit ve yaygın olarak kullanılan yöntem k-Ortalamlar algoritmasıdır (Rokach ve Maimon, 2010). Bu algoritma veriyi merkezleri veya ortalamalarıyla gösterilen k sayıda kümeye $(C_1, C_2, \dots, C_k)$ bölüştürür. Her kümenin merkezi o kümeye ait tüm örneklerin ortalaması olarak hesaplanır.

Algoritma, rasgele seçilen veya bazı sezgisel prosedürlerle seçilen ilk küme merkezleriyle çalışmaya başlar. Her basamakta her bir örnek diğer bir örnekle arasındaki Öklid mesafesine bakılarak belirlenen en yakın küme merkezine atanır. Daha sonra küme merkezleri aşağıdaki gibi hesaplanır. Burada N_k , k kümesine ait örnek sayısını, μ_k ise k kümesinin ortalamasını belirtir. Her bir kümenin merkezi o kümeye ait örneklerin ortalamasıdır.

$$\mu_k = \frac{1}{N_k} \sum_{q=1}^{N_k} x_q$$

k-ortalamlar algoritmasının popüler olmasındaki nedenlerden biri doğrusal karmaşıklığıdır. Örnek büyüklüğü çok olsa bile, bu algoritma hesaplama açısından etkilidir. Bu sayede k-ortalamlar algoritması diğer kümeleme algoritmalarıyla karşılaştırıldığında avantaja sahiptir.

4.2.2. Sınıflandırma

Analizden sonra sonuçlar hakkında daha net görüşe ve genel bir çıktıya sahip olmak gerekmektedir. Grupların olması tarif etmede açıklayıcı olacaktır. Bu da SPSS’te aralıklar ve istenilen grup sayılarına göre gruplandırma ile yapılabilir, fakat sınıflandırma algoritmaları daha anlamlı kategoriler vermektedir. Bu yüzden bu çalışmada yapay sinir ağları bayes ağları ve C5.0 gibi sınıflandırma algoritmaları kullanılmıştır.

Bayes Ağ Modeli

Bir Bayes ağı, değişkenleri (sıkça düğüm olarak ifade edilir) bir veri kümesi içinde gösteren ve bu kümeler arasındaki olasılıklı veya şartlı bağımsızlıkları gösteren grafiksel bir modeldir. Bir SPSS uygulaması olarak Bayes ağları kategorik çıktıları tahmin etmede kullanılır (Di Marco ve Hanlon, 2003).

Ağ, Bayezyen olasılık teorisine dayanır. Bu teori, özellikle hedef değişkenin olmak üzere, ilgilenilen değişkenlerin ortak dağılımlarını ve her bir değişkenin en üstün dağılımlarını kullanır.

Değişken seçimi değişken kümelerini azaltmak için yapılabilir çünkü Bayes ağları en iyi şekilde küçük tahmin gruplarında çalışır.

Bayes ağları iki temel parçadan oluşmaktadır (Tobergte ve Curtis, 2013). İlki, düğümlerin tahmin değişkenlerini ifade ettiği ve arkların stokastik bağımlılık ilişkilerini ifade ettiği döngüsel olmayan bir grafiktir. Özellikle, ağdaki a_j düğümüyle ilişkili X_j değişkeninin, a_j 'den önceki düğümlerle ilişkili değişkenlere bağımlı olduğu ve a_j 'den doğrudan ulaşılabilen düğümlerle ilişkili değişkenlere şartlı olarak bağımsız olduğu varsayılmaktadır.

Bayes ağlarının ikinci parçası da her bir değişken için atanmış şartlı olasılıkların tablosunu içerir. Özellikle, X_j değişkeni ile ilişkili tablo, C_j 'nin a_j düğümünden önceki düğümlerle ilişkili açıklayıcı değişkenlerinin kümesini açıkladığı durumda, $P(X_j | C_j)$ şartlı olasılığını içerir. Bu da veri kümesindeki yakın tekrarlamalara dayanarak tahmin edilir.

Şartlı olasılıkları tahmin etmek için öngörücü değerlerin tüm olası kombinasyonlarının hesaplanmasında gerekli karmaşıklık Bayes ağlarında kısıtlıdır. Çünkü hesaplama, ağda bulunan önceki bağlantılarla belirlenen şart ilişkilerini azaltır.

Bayes ağları ve Bayes metotları aşağıdaki dört faydayı sağlamaktadır (Heckerman, 1996);

Bayes ağları tamamlanmamış veri kümelerinin üstesinden gelebilir. Girdilerden biri gözlenmemişse modellerin çoğu ancak hatalı tahmin yapabilmektedir. Çünkü onlar giriş değişkenleri arasında korelasyonlar tanımlamamaktadır. Bayes ağları bu bağımlılıkları tanımlamak için doğal bir yol sunar.

Bayes ağları nedensel ilişkiler kurarak öğrenmeye izin verir.

Bayes istatistiksel teknikleriyle birlikte çalışan Bayes ağları çevrimin bilgi ve verilerinin kombinasyonuna olanak tanır. Gerçek hayatta analiz gerçekleştirenler öncelikli olanın veya çevrim bilgisinin önemli olduğunu bilmektedir (özellikle de veri kusurlu veya pahalı ise). bazı ticari sistemlerin (uzman sistemler gibi) tek başına öncelikli olan bilgidir.

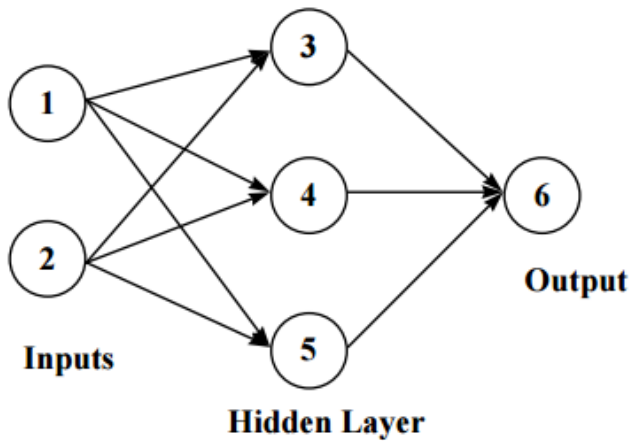
kurulabilmesi, öncelikli bilginin gücünün bir göstergesidir. Bayes ağlarının, nedensel öncelikli bilginin tanımlanmasını apaçık bir şekilde yapan nedensel bir anlamı vardır. İlave olarak, Bayes ağları nedensel ilişkilerin gücünü olasılıklarla tanımlar. Sırasıyla öncelikli bilgi ve veri Bayesyen istatistikteki iyi çalışılan tekniklerle kombine edilebilir.

Bayes ağları ve diğer model tipleriyle çalışan Bayes yöntemleri verinin aşırı uyum göstermesini önlemek için etkili ve temel bir yaklaşım sağlar. Test etmek için uygun bazı verilerin tutulmasına gerek yoktur. Bayes yaklaşımı kullanılarak modeller, tüm uygun verilerin öğrenme için kullanılabilirdiği bir şekilde düzgünleştirilebilir.

Sinir Ağları Modeli

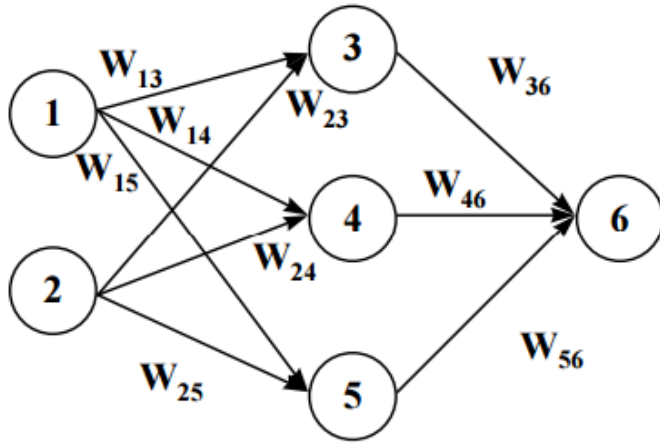
Sinir ağları nöronlardan oluşan biyolojik sistemlerin davranışını simüle etmeye yarar. En basit modellerin sunulduğu 1950'lerden beri, sadece sınıflandırma için değil sürekli hedef olayların regresyonunda da öngörücü amaçlarda kullanılmıştır (Tobergte ve Curtis, 2013).

Bir sinir ağı, biyolojik analogide nöronları temsil eden ve arklarla bağlanan düğümlerin, dendrit ve sinapslarla ilişkilendirilmesine yönelik bir grafiktir. Her ark bir ağırlıkla ilişkilendirilmektir. Her bir düğümde gelen arklardaki düğümlerle girdi olarak alınan değerlere uygulanan bir aktivasyon fonksiyonu tanımlanmaktadır ve arkların ağırlıklarıyla düzeltmeler yapılmaktadır. Öğrenme aşaması, öğrenme kümesinde bulunan gözlemlerin arka arkaya sırası analiz edilerek ve arklara ilişkin ağırlıkların her işlemde modifiye edilmesi ile gerçekleştirilir.



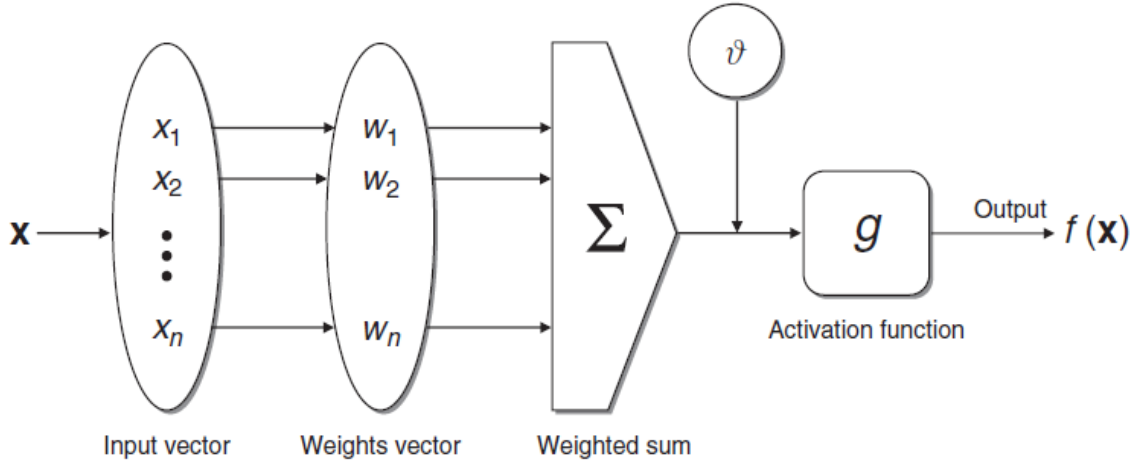
Şekil 4.3. Gizli katmanlı bir sinir ağı

Bir sinir ağı (Şekil 4.3) her bir düğümün öngörücü bir değişkenle ilişkili olduğu bir giriş katmanıyla başlar (Edelstein, 1999). Bu giriş düğümleri, gizli bir katmandaki belli sayıdaki düğümlere bağlanır. Gizli katmandaki bu düğümler başka bir gizli katmandaki düğümlere veya çıkış katmanına bağlanabilir. Çıkış katmanı bir veya daha fazla yanıt değişkenini içerir. Giriş katmanından sonra, her bir düğüm bir giriş kümesi alır, onları bağlantı ağırlığıyla çarpar (Şekil 4.4), birlikte ekler, aktivasyon fonksiyonu olarak bilinen bir fonksiyon uygular ve çıktıları sonraki katmandaki düğümlere aktarır.



Şekil 4.4. Ağırlıklandırılmış sinir ağı (W_{xy} x düğümünden y düğümüne kadar olan ağırlık)

Sinir ağının en basit hali Şekil 4.5'te gösterilmiştir ve gelen bağlantılar boyunca giriş değerini $(x_1, x_2, \dots, x_n)$ alan ve $f(x)$ çıktı değeri olarak geri dönen bir nöronu ifade etmektedir (Tobergte ve Curtis, 2013). Giriş değerleri açıklayıcı özelliklerin değerleriyle çakışır. Çıkış değeri de y yanıt değişkeninin tahminini yapar. n giriş bağlantılarından her biri w_j ağırlığı ile ilişkilidir. Bir g aktivasyon fonksiyonu, çarpıklık katsayısı θ de atanır.



Şekil 4.5. Sinir ağında bulunan tek bir birimdeki işlemler

Geri yayılım algoritması olarak bilinen, düğümlerdeki tüm ark ve çarpıklıkların ağırlıklarını bulan bir yöntem, tek bir algılayıcı için tanımlanan algorithmadan çok da farklı olmayan bir mantık izler. Ağırlıklar keyfî seçilmiş bir yoldan, örneğin; rasgele üretilmiş sayılara eşit bir değer belirlenerek başlatılır. Bu yüzden öğrenme kümesinin örnekleri, tahminin ve yanlış sınıflandırmaların hesaplanması için ağırlıkların mevcut değerleri her bir işlemde sırayla kullanılmasıyla incelenir. Yanlış sınıflandırmalar ağırlık değerlerinin doğrulanması için kullanılır. Sonraki bir zamanda prosedürdeki izleyen örnekleri analiz etmede kullanılır. Ağırlıklar, değişim ölçüsü yöntemlerinin bir çeşidi olan bir algoritma ile güncellenir.

Sinir ağlarının güçlü yönlerinden birisi hem sınıflandırma hem de regresyon problemlerine uygulanabilen bir öğrenme mekanizması olmalarıdır. İlişkili olmayan veya gereksiz değişkenler işe yaramayan değerlerin katsayılarına bakılarak analizden çıkarılabildiğinden özellik seçiminde otomatik olarak yer almaktadır. Ancak, sinir ağları model öğrenmesi için çok uzun zaman almaktadır ve örneklerin analiz edildiği sıraya bağlı yorumlanabilir sonuçlar sunmaktadır. Ayrıca gürültü tarafından etkilenen veriyle uyumlu olarak daha az dayanıklılık gösterir.

C5.0 Algoritması

Karar ağaçları, ayrılan gruplar arasındaki mesafeyi maksimize etmek amaçlandığında verinin kesikli gruplara ayrılmasıyla üretilir. Kategorik değişkenleri tahmin etmede

kullanılan karar ağaçları, örnekleri yerleştirmelerinden dolayı sınıflandırma ağaçları olarak tanımlanır (Edelstein, 1999). En popüler sınıflandırma ağaçları C4.5, C5.0 ve C&RT'dir.

C5.0 algoritması hem ID3 hem de C4.5 algoritmalarının genişletilmiş halidir (Neelamegam ve Ramaraj, 2013). Büyük veri kümelerine uygulanabilen bir sınıflandırma algoritmasıdır. Hız, hafıza ve verim açısından C4.5'ten daha iyi sonuçlar vermektedir. C5.0 modeli maksimum bilgi kazancını (information gain) veren bir alana dayanarak, örnek kümesini bölmesiyle çalışır. C5.0 modeli en büyük bilgi kazanç alanına sahip örnekleri bölebilir.

Bilgi kazancı, bir özelliğin hedef konsepte göre öğrenme verisini ne kadar iyi ayıracağına karar vermek için kullanılır (Estabrooks, 2000). Entropi olarak bilinen ve bilgi teorisiyle ortak olarak kullanılan bir ölçüye dayanır. Entropi, bir S kümesinin rasgele seçilmiş bir üyesi sınıflandırılırken kullanılması gereken minimum parça sayısıdır. Boolean hedef konseptine göre S 'nin entropisi aşağıdaki gibi tanımlanmıştır.

$$Entropi(S) = -p_{(+)} \log_2 p_{(+)} - p_{(-)} \log_2 p_{(-)}$$

Burada $p_{(+)}S$ kümesindeki pozitif örneklerin, $p_{(-)}$ negatif örneklerin oranıdır.

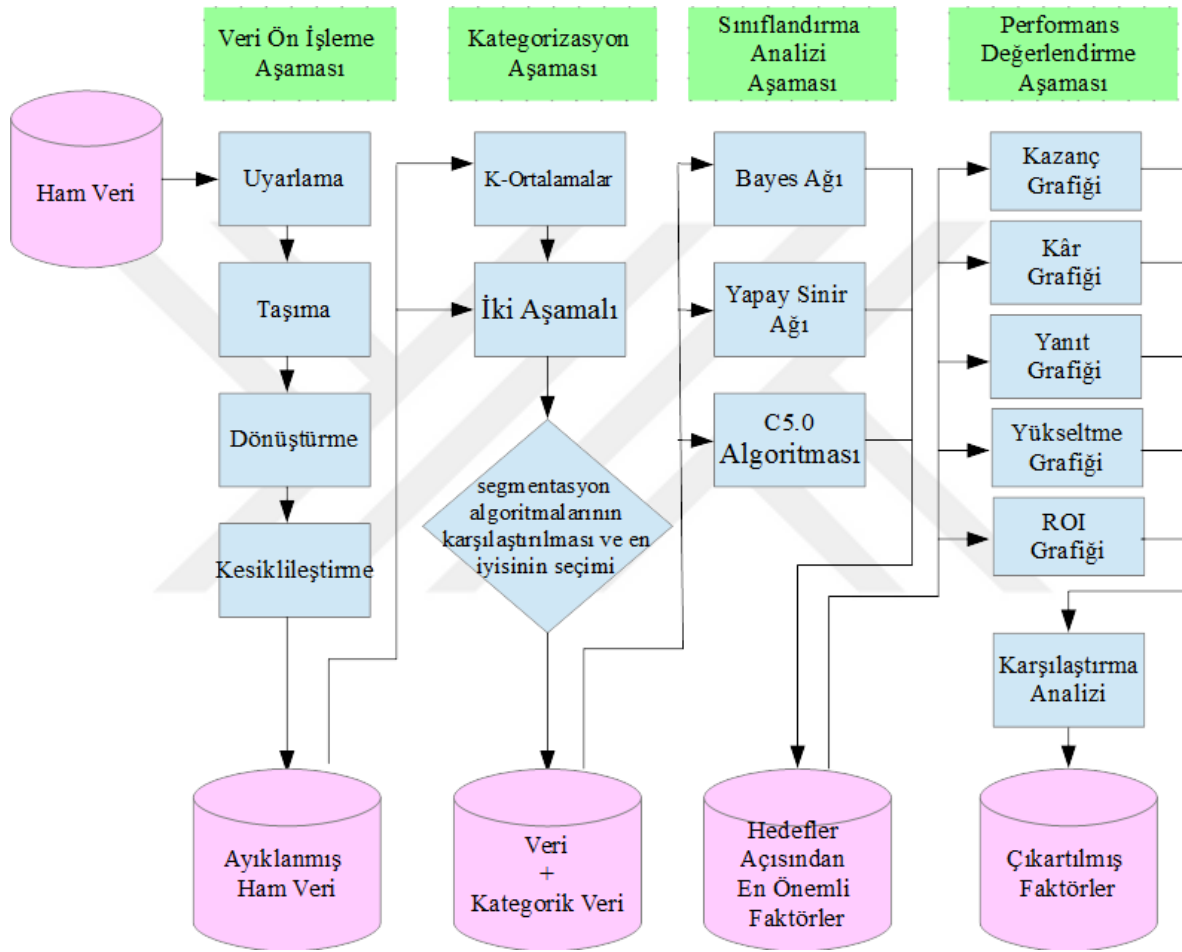
$$Kazanç(S, A) = Entropi(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropi(S_v)$$

Kazanç hesaplanırken kullanılan A değerleri, tüm A özelliklerini temsil etmektedir. S_v , A özelliği için v değerini alan S kümesindeki örneklerin kümesidir.

Önceki ayırmadan elde edilen örnek alt küme sonradan tekrar ayrılacaktır (Neelamegam ve Ramaraj, 2013). Örnek alt küme ayıramayacak hale gelene kadar bu işlem tekrarlanır. Sonunda, en alt seviyedeki ayırım incelenecek ve modele dikkate değer katkıda bulunmayan alt kümeler reddedilecektir. C5.0, veri kümesinden elde edilen çok değerli ve eksik özelliklerle de çalışabilmektedir.

4.3. Çalışmada Kullanılan Metodoloji

Bu çalışmada, Araç Sertifikasyon Ajansı'ndan ve Mercedes Benz internet sayfasından alınan ham veri kullanılmıştır. Bu veri üzerindeki uygulamaların basamakları CRISP DM prosedürü ve Sarvari, Ustundag, ve Takci (2016) tarafından önerilen bir metodolojinin adımları dikkate alınarak Şekil 4.6'daki gibi tasarlanmış ve uygulanmıştır.



Şekil 4.6. Uygulanan metodolojinin işlem adımları



5. UYGULAMA

5.1. Veri

Sözlük ve kitaplarda verinin çeşitli tanımları vardır, fakat esas olarak bu tanımların çoğunun anlattığı şey verinin kaydedilmiş vakalar, olaylar, işlemler ve benzeri olduğudur (Gordon, 2007). Verinin kaydedilen vakaların kümesi olduğu birçok programlama kitabında bulunmaktadır.

5.1.1. Veri üretimi

İnternette veri toplamak veri üretiminde bir süreçtir. Html, xml, vb. formattaki internet sayfaları kod, stil, kodlama uygulamalarındaki istisnalar ve ihlallerin çeşitliliğinden dolayı biçimlendirilmemiş veri kaynakları olarak düşünülür. Bu çeşitlilikten dolayı, internette veri çekme elde edilmek istenilen bilginin kaynağına dayanan büyük ölçüde kişiselleştirilebilir bir süreçtir. Veri çekmenin tanımı, verinin biçimlendirilmemiş şeklini alarak bu bilgiyi biçimlendirilmiş veri kümesine dönüştürmektir (Gottlob, 2016).

Bu çalışmada, Araç Sertifikasyon Ajansı (Vehicle Certification Agency-VCA)’ndan alınan 2015 yılında kaydedilmiş ve kullanılmış otomobillerin yakıt tüketimi ve emisyon verileri (Band, 2016) ve Mercedes Benz’in araç spesifikasyon verileri (“Technical Specification”, 2004) kullanılmıştır.

5.1.2. Betimleyici istatistikler

Başlangıç veri kümesi farklı otomobil modellerini temsil eden 2667 gözlem ve 27 değişken içermekteydi ve bütün veriler Excel’in bir sütununda yer almaktaydı. Bu yüzden her bir değişken SPSS’te kullanılabilecek şekilde sütunlara ayrıldı.

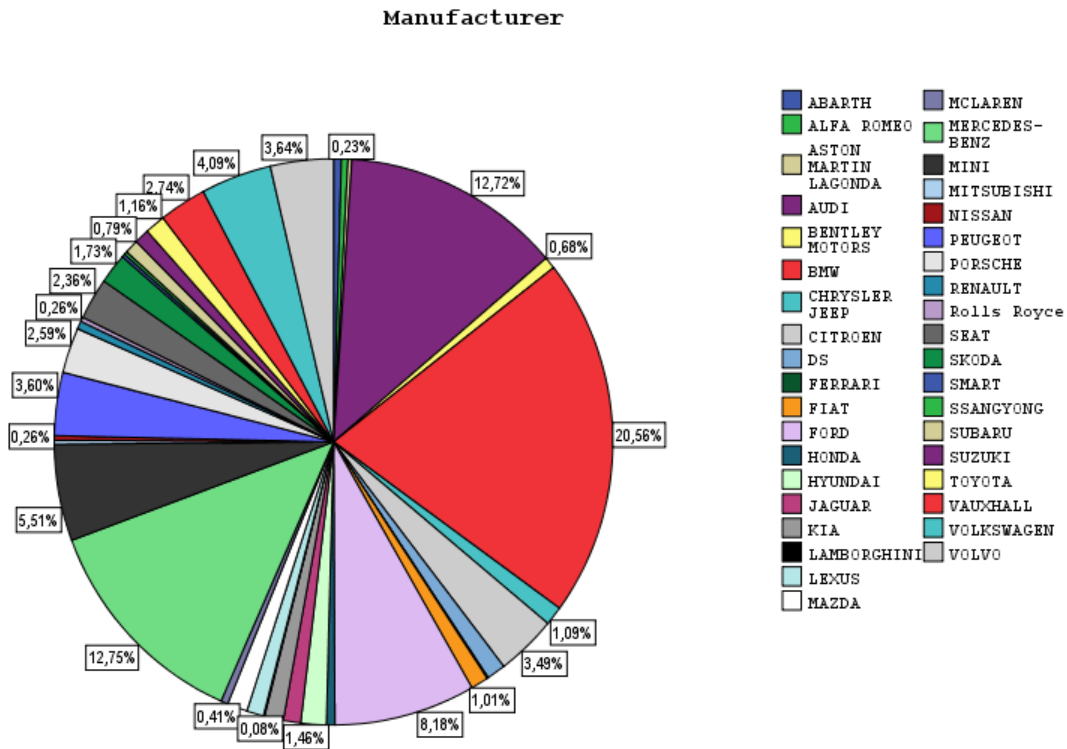
Değişkenler; otomobil üreticisinin adı, aracın modeli, araçla ilgili detaylı bilgiler, şanzıman tipi (vites kutusu), motor kapasitesi, yakıt tipi, elektrik enerjisi kullanımı(mil/kW saat), (elektrikli araçlar için) bir şarj ile gidilebilecek maksimum mesafe (km ve mil), şehir içi çeşitli ortamlardaki yakıt tüketimi (litre ve galon cinsinden), üretilen karbondioksit miktarı, araç ağırlığı, 12000 mildeki yakıt maliyeti, elektrik maliyeti, toplam maliyet, Euro standardı,

gürültü seviyesi, nitrojen oksitleri emisyonu, hidrokarbon emisyonu, partikül emisyonu gibi parametreleri içermektedir.

Üretici firma

Mevcut üreticiler, Avrupa M1 standardına göre otomobil üreten şirketlerdir. Büyük hacimli veya 4 tekerlekli L tipi araçlar üretmemektedir (Band, 2016).

Kullanılan veri kümesinde 38 üretici firmaya ait çeşitli özel araç modellerine dair bilgi bulunmaktadır (Şekil 5.1).



Şekil 5.1. Üreticilerin dağılımı

Bu araçların çoğu, Çizelge 5.1’de görüldüğü gibi BMW (%20.6), Mercedes (%12.8), Audi (%12,7) markalarına aittir. İlgili histogram EK-1’de verilmiştir.

Avrupa Pazar Araştırmalarına göre 2013’teki en geniş pazar %25’lik pazar payı ile Almanya’ya aittir (ICCT Europe, 2013). Benzer bir yorum, bu çalışmadaki 2016 veri kümesinden çıkan istatistiksel sonuçlarla da yapılabilir. Bunlar, Alman üreticilerin 2020 ya da 2030 hedeflerine yönelik analizlerde ayrıca kullanılabilir.

Çizelge 5.1. Üreticilerin dağılımı

		Sıklık	Oran	Geçerli Oran	Kümülatif Oran
Valid	ABARTH	12	,5	,5	,5
	ALFA ROMEO	10	,4	,4	,8
	ASTON MARTIN	6	,2	,2	1,1
	LAGONDA				
	AUDI	339	12,7	12,7	13,8
	BENTLEY MOTORS	18	,7	,7	14,4
	BMW	548	20,6	20,6	35,0
	CHRYSLER JEEP	29	1,1	1,1	36,1
	CITROEN	93	3,5	3,5	39,6
	DS	29	1,1	1,1	40,7
	FERRARI	2	,1	,1	40,7
	FIAT	27	1,0	1,0	41,7
	FORD	218	8,2	8,2	49,9
	HONDA	13	,5	,5	50,4
	HYUNDAI	39	1,5	1,5	51,9
	JAGUAR	27	1,0	1,0	52,9
	KIA	29	1,1	1,1	54,0
	LAMBORGHINI	2	,1	,1	54,1
	LEXUS	26	1,0	1,0	55,0
	MAZDA	31	1,2	1,2	56,2
	MCLAREN	11	,4	,4	56,6
	MERCEDES-BENZ	340	12,8	12,8	69,4
	MINI	147	5,5	5,5	74,9
	MITSUBISHI	6	,2	,2	75,1
	NISSAN	7	,3	,3	75,4
	PEUGEOT	96	3,6	3,6	79,0
	PORSCHE	69	2,6	2,6	81,5
	RENAULT	12	,5	,5	82,0
	Rolls Royce	7	,3	,3	82,3
	SEAT	63	2,4	2,4	84,6
	SKODA	46	1,7	1,7	86,3
	SMART	5	,2	,2	86,5
	SSANGYONG	5	,2	,2	86,7
	SUBARU	21	,8	,8	87,5
	SUZUKI	23	,9	,9	88,4
	TOYOTA	31	1,2	1,2	89,5
	VAUXHALL	73	2,7	2,7	92,3
	VOLKSWAGEN	109	4,1	4,1	96,4
	VOLVO	97	3,6	3,6	100,0
	Total	2666	100,0	100,0	

Şanzıman

Üreticiler vites kutusunu tarif edebilmek için çeşitli kısaltmalar kullanmaktadır. Tam listesi olmamasına rağmen veri kümesindeki araçlar aşağıdaki vites tiplerini içermektedir (Band, 2016):

A - otomatik

A4 - otomatik 4 hız

4Atx2 - otomatik 4 hız, yüksek ve düşük aralıkta vites

AV - AV otomatik değişken

A5 – otomatik 5 hız

A7 - otomatik 7 hız

A/SAT5 - otomatik / yarı otomatik şanzıman 5 hız

ASM - otomatik vites manuel

AWD – tüm teker sürüşü

CVT – sürekli değişken şanzıman

D – dizel motorlu araçlar

D6 – direkt vites 6 hız

DCT7 – çift debriyajlı şanzıman 7 hız

M - manuel

M5 - manuel 5 hız

5MTx2 - manuel 5 hız, yüksek ve düşük aralıklı vites

MTA – otomatik değişmenin olduğu manuel şanzıman

M6 - manuel 6 hız

M6/S6 - manuel 6 hız / sıralı 6 hız

MULTI5 – çok modlu 5 hız

QA5 – 4 tekerli sürüş, otomatik 5 hız

QA6 - 4 tekerli sürüş, otomatik 6 hız

QD6 - quattro direkt vites 6 hız

QM5 - 4 tekerli sürüş, manuel 5 hız

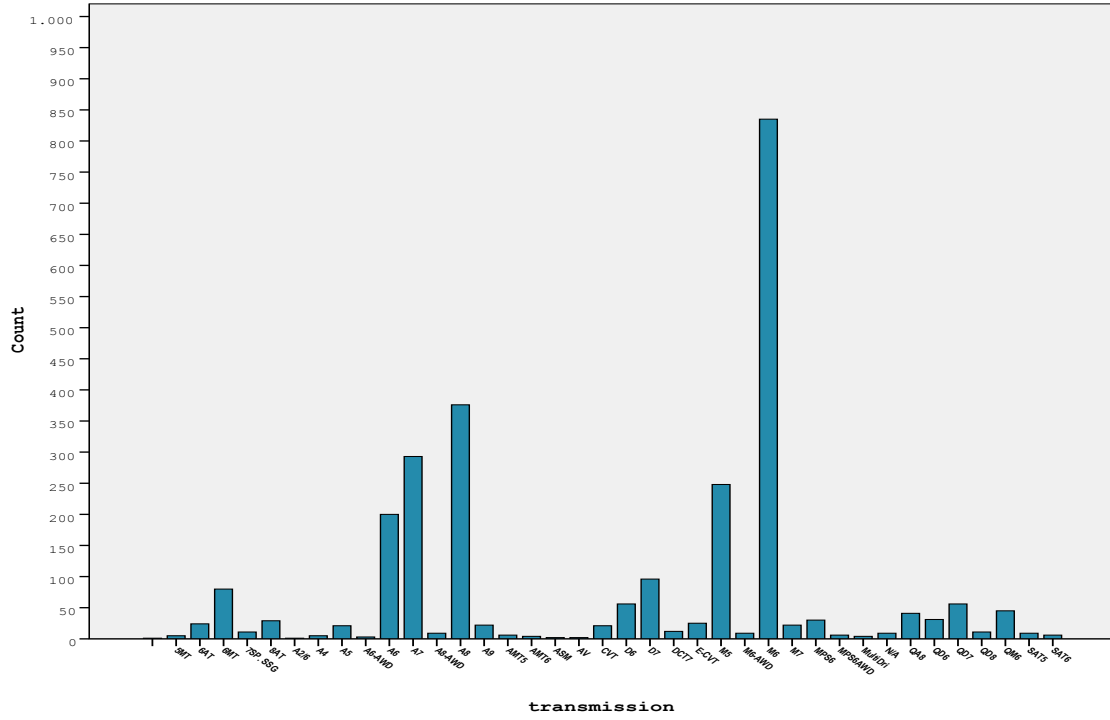
QM6 - 4 tekerli sürüş, manuel 6 hız

S/A6 – yarı otomatik şanzıman 6 hız

SAT5 - yarı otomatik şanzıman 5 hız

SAT6 - yarı otomatik şanzıman 6 hız

SMG7 – sıralı manuel vites kutusu 7 hız



Şekil 5.2. Şanzıman tipleri

Şekil 5.2’de görüldüğü gibi veri kümesinde 37 tipte şanzıman kullanılmaktadır ve araçların çoğu manuel 6 hız (M6) vites kutusuna sahiptir.

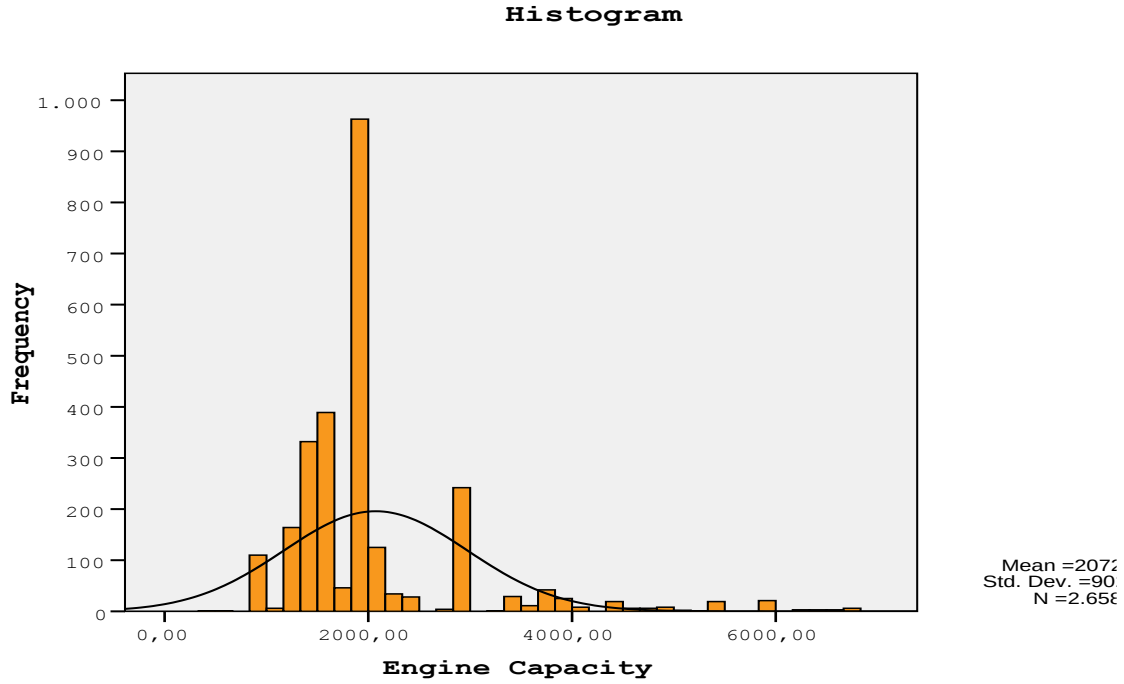
Motor kapasitesi

Motor kapasitesi santimetre küp cinsinden hesaplanmıştır. Veri kümesinde özel araçların motor kapasitesi 395 ile 6752 cm³ arasında değişmektedir ve ortalama 2072 cm³’tür (Çizelge 5.2).

Çizelge 5.2. Motor kapasitesinin istatistikleri

N	Valid	2658
	Missing	8
Mean		2072,3849
Std. Error of Mean		17,50498
Median		1984,0000
Std. Deviation		902,48312
Variance		814475,778
Skewness		2,288
Std. Error of Skewness		,047
Kurtosis		6,695
Std. Error of Kurtosis		,095
Range		6357,00
Minimum		395,00
Maximum		6752,00
Sum		5508399,00
Percentiles	25	1560,0000
	50	1984,0000
	75	1999,0000

Araçların çoğu 1984 cm³'lük motor kapasitesine sahiptir ve bu, ortalamanın altında olduğu için Çizelge 5.2 ve Şekil 5.3'teki gibi sağa doğru pozitif çarpıklık ve basıklık göstermektedir. Bunun sebebi Bentley, Rolls Royce, Lamborghini gibi lüks otomobillerin üst sınır değerini yükseltmesidir (Bkz. EK-3).



Şekil 5.3. Motor kapasitesi dağılımı

Yakıt tipi

Bu çalışmada kullanılan veri kümesi alternatif yakıtları kullanan belli bir sayıda araç içermesine rağmen genellikle petrol ve dizel kullanan araçlar bulunmaktadır. Başlangıç verisinde bulunan araçlar aşağıdaki yakıt tiplerini kullanmaktadır:

Elektrik: Bataryalarda depolanmış elektrik enerjisini kullanan bir veya daha fazla elektrikli motor kullanan araçlar mevcuttur.

Hibrit-elektrik, petrol-elektrik veya dizel-elektrik: Geleneksel yakıtlarla birlikte depolanan elektrik enerjisini de kullanan araçlar mevcuttur.

Şarj edilebilir elektrik hibrit: Depolanan elektrik ve bir içten yanmalı motorla beraber çalışabilen araçlar mevcuttur.

Aküler: Motordan ve dinamolarla elektrik depolayabilen veya bir dış kaynaktan üretilen elektrikle çalışan bataryalar kullanılmaktadır.

CNG (car that runs on Compressed Natural Gas): Sıkıştırılmış doğal gazla çalışan araçlar mevcuttur.

LPG (car that runs on Liquefied Petroleum Gas): Yoğunlaştırılmış petrol gazıyla çalışan araçlar mevcuttur.

Çizelge 5.3. Yakıt tiplerinin dağılımı

	Sıklık	Oran	Geçerli Oran	Kümülatif Oran
Valid	1	,0	,0	,0
Diesel	1294	48,5	48,5	48,6
Diesel E	14	,5	,5	49,1
Electric	19	,7	,7	49,8
Petrol	1303	48,9	48,9	98,7
Petrol E	10	,4	,4	99,1
Petrol H	25	,9	,9	100,0
Total	2666	100,0	100,0	

Bu çalışmada kullanılan veri kümesi sadece dizel, sadece petrol, sadece elektrik, elektrikli dizel, elektrikli petrol ve petrol hibrit yakıt tiplerini kullanmaktadır. Çizelge 5.3'teki gibi, araçların büyük çoğunluğu Avrupa Pazar Araştırmaları (ICCT Europe, 2013) ile benzer şekilde yakıt olarak yalnızca dizel (%48,5) veya yalnızca petrol (%48,9) kullanmaktadır. Bundan dolayı analizde diğer yakıt tipleri dikkate alınmamıştır.

Yakıt tüketimi

Başlangıç verisindeki yakıt tüketim miktarları, çeşitli test çevrimlerinin sonuçları olarak elde edilmiştir. Bu testlerle ilgili detaylı bilgiler tezin uygulama kısmında verilmektedir. Ayrıca

litre cinsindeki çevrimlerin İngiliz ölçü birimlerine göre mili galon (mpg) şeklinde dönüşümleri de mevcuttur.

Petrol ve dizel kullanan özel araçların dağılımları EK-4'te gösterilmiştir. Histogramlara bakılarak yakıt tüketim şekillerinin normal dağılım gösterdiği söylenebilir.

Maksimum mesafe

Maksimum mesafe, tamamen şarj edilen bir elektrikli aracın kilometre cinsinden gidebileceği maksimum mesafeyi ifade etmektedir. Bu parametre, veri kümesinde bulunan ve elektrik enerjisi kullanan sadece 19 adet araç için hesaplanabilmektedir. Maksimum mesafe 31 ile 212 kilometre arasında değişmekte ve her şarjda ortalama 103 km olarak hesaplanmaktadır.

İstatistiklere bakıldığında, analizde kullanılacak veri kümesi bazı düzenlemeler yapılarak hazırlanmıştır.

- Her bir örneğin açıklamasında araca dair gizlenmiş birçok bilgi bulunmaktadır. Bunlar aracın serisi, tekerlek çapı ve motor tipi şeklinde üç farklı değişken olarak sütunlara yerleştirilmiştir.
- Farklı ölçü birimiyle gösterilen aynı yakıt tüketimlerinin galon cinsinden olanları veri kümesinden çıkartılmıştır.
- Veri kümesinde sadece 19 elektrikli araç olmasından dolayı elektrik enerjisi kullanımları ve maksimum mesafe de veri kümesinden çıkartılmıştır. Sadece petrol ve sadece dizel kullanan araçlar dikkate alınmıştır.
- Analizde, veri kümesindeki örneklerin çoğunluğunu oluşturmalarından ve tüm verileri temsil etmede yeterli olduklarından Alman BMW, Mercedes ve Audi markalarına ait araçlar dikkate alınmıştır.
- Araçların ağırlığı, başlangıçta SPSS'te beş kategoriye ayrılmış fakat daha güvenilir sonuçlar elde etmek için segmentasyon algoritmalarıyla elde edilen kategorilerin kullanılması daha uygun görülmüştür.

5.2. Tasarım Aşaması

Euro 6 verisi analizde kullanıma elverişli değildi ve işlenmesi gerekmektedir. Bunun için, 4.kısımda belirtilen metodolojinin (Şekil 4.6) veri işleme adımları takip edilmiştir. Teknik özellik verisinde yeni değişkenler ve eksik veriler eklenmiş, istatistiklere göre analizi bozacak uç ve süreksiz veriler silinmiştir. Veri kümesi analizde kullanılacak şekle dönüştürülmüş, sürekli değişkenler k-ortalamalar ve iki aşamalı kümeleme algoritmasıyla kategorilere ayrılmıştır. Düzenlenen veri kümesine göre segmentasyon teknikleri karşılaştırılmış, daha iyi olan seçilmiştir. Kategorik veri, başlangıç verisiyle birlikte kaydedilerek sınıflandırmada kullanılmıştır.

Analiz aşamasında, sinir ağı, Bayes ağı ve C5.0 algoritması gibi teknikler kullanılmış, sınıflandırma yapılarak belirlenen hedefleri (yakıt tüketimi, emisyon, gürültü seviyesi) etkileyen en önemli faktörler bulunmuştur. Her teknik ve bulunan faktörler çeşitli grafikler (fayda, kazanç, yükseltme, yanıt ve ROI) yardımıyla görselleştirilmiştir. Performans değerlendirme aşamasında ise, bu grafiklerin sonuçları birbiriyle karşılaştırılarak sonuçlar özetlenmiştir. Bulunan sonuçlar ve önemli faktörler kaydedilmiştir.

5.3. Analiz Aşaması

Yakıt tüketimi test prosedürleri resmi olarak 1970'lerden beri kullanımdadır. Avrupa Birliği Direktifi (EU Directive) 80/1268/EEC'nin 692/2008 sayılı düzenlemesi Euro 5 standardından itibaren 1 Ocak 2001'den itibaren satışa sunulan tüm yeni otomobillerin yaptırması gereken testleri açıklamıştır (Offices et al., 2016). Konvansiyonel içten yanmalı motorla çalışan araçlar için mevcut test şehir içi, şehir içi ekstra ve şehir içi kombine olmak üzere üç kısımdan oluşmaktadır. Test çevrimi, hava kalitesini bozan (kirleten) emisyon için kullanılan ile aynıdır. Test edilen arabalar ısınmış olmalıdır, bu yüzden test öncesinde en az 3000 kilometre sürülmüş olmalıdır.

Emisyon seviyelerini değerlendirmek için farklı bir test ortamı gereklidir. Bu sebeple, testlerin çoğu Yeni Avrupa Sürüş Çevrimi (the New European Driving Cycle-NEDC) ne yakındır. Bu çevrimde, mevcut emisyon testi onaylanacak araçlar içindir ve buna göre stratejiler tasarlanmaktadır. Test programı bu çevrimin varyasyonları etrafında ve emisyon

laboratuvarlarında şekillendirilir. Diğer testler de kamusal yollarda araçların gerçek hayattaki emisyon performansını ölçmek için yapılmaktadır (Şekil 5.4).



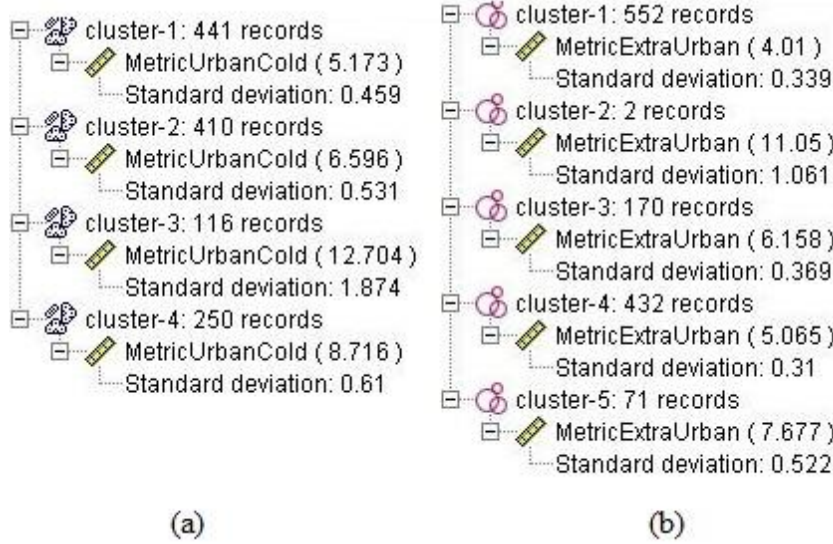
Şekil 5.4. Her test aracına kurulan portatif emisyon ölçüm sistemi (PEMS) (Offices ve diğerleri, 2016)

Birleşmiş Milletler günümüzde the Worldwide Harmonized Light Vehicles Test Procedure (WLTP) adında yeni bir test prosedürü geliştirmektedir. Bu prosedür tamamlanmak üzeredir ve mevcut araç testlerinden bilinen bazı konulara işaret edecektir. WLTP'nin 2016/2017'de Avrupa Birliği'nde uygulanması beklenmektedir.

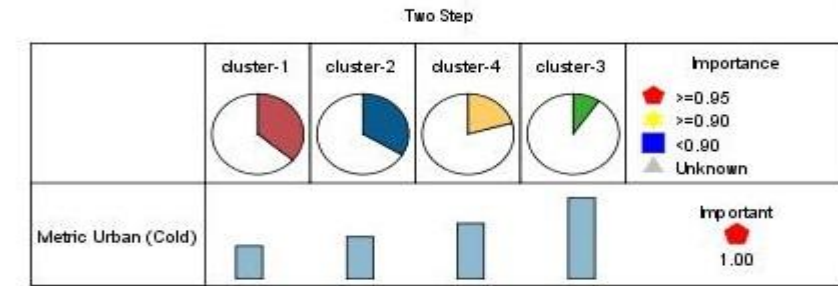
5.3.1. Şehir içi yakıt tüketimi

Şehir içi yakıt tüketim testi 20°C-30°C sıcaklıktaki bir laboratuvar ortamında soğuk başlatıldıktan sonra (motorun birkaç saat çalıştırılmamasının ardından) dönen bir yolda gerçekleştirilmektedir (Offices ve diğerleri, 2016). Test çevrimi, bir dizi hızlanmalar, sabit hızlar, yavaşlamalar ve boşta çalışmaları içerir. Maksimum hız saatte 50 kilometre, ortalama hız saatte 19 kilometre ve toplam mesafe 4 kilometredir. Veri kümesinde bulunan değerler 100 kilometredeki yakıt tüketimini litre cinsinden ifade etmektedir.

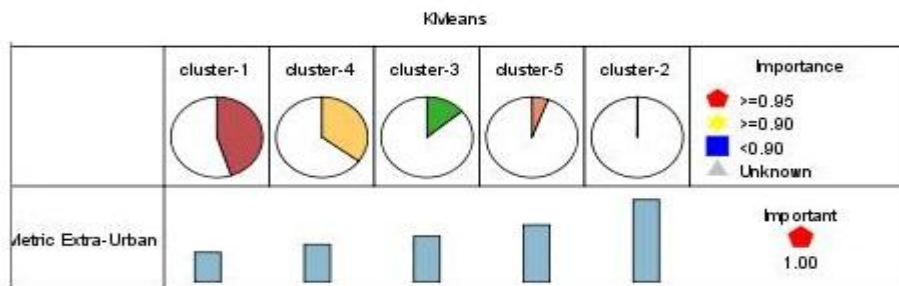
Şehir içi yakıt tüketimi değerleri (metric urban cold) k-ortalamlar ve iki aşamalı kümeleme algoritması kullanılarak Şekil 5.5'teki gibi gruplandırılmıştır. Her iki yöntemde de 5 kategori bulunmasına rağmen bu grupların merkez değerleri farklıdır. k-ortalamlar ile bulunan cluster-2'de sadece 1 örnek bulunduğu için analizde diğer 4 grup dikkate alınmıştır.



Şekil 5.5. Şehir içi yakıt tüketimi için (a) iki aşamalı kümeleme algoritması ve (b) k-ortalamlar algoritması ile belirlenen kümeler



(a)



(b)

Şekil 5.6. Şehir içi (a) iki aşamalı kümeleme algoritması ve (b) k-ortalamlar algoritması ile bulunan kümelerin dağılımlarının karşılaştırılması

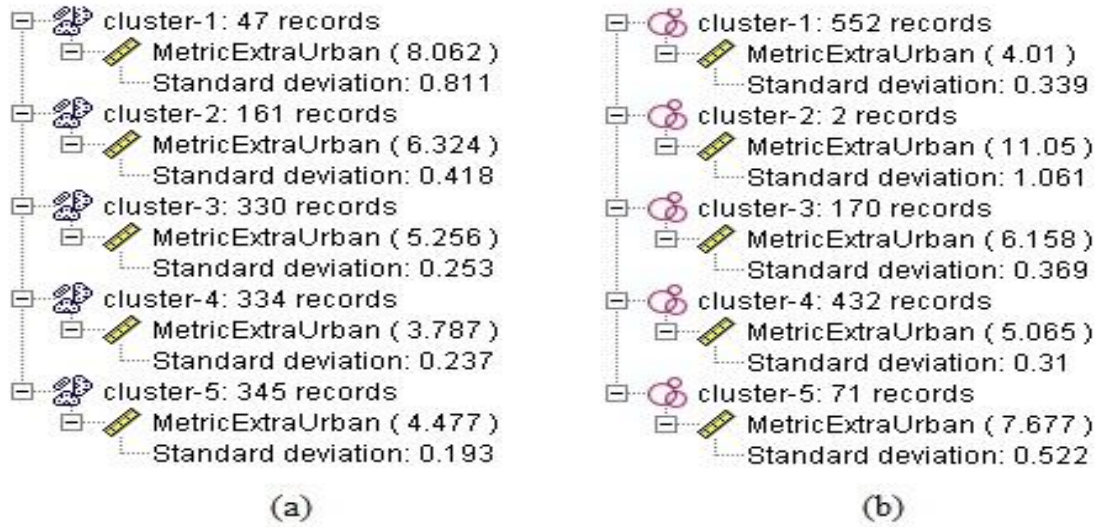
Şekil 5.5a ve Şekil 5.6a'daki gibi iki aşamalı kümeleme kullanıldığında örneklerin çoğu (475 adet) cluster-3'e dahil olmuştur. Küme ortalamaları sırasıyla 14.25 l, 11.11 l, 5.26 l, 6.68 l ve 8.65 l olmuştur. k-ortalamalar ile bulunan kategorilerden en kalabalığı (cluster-1) 718 örnek içermektedir (Şekil 5.5b ve 5.6b). Diğer kategoriler sırasıyla 20.5 l, 10.88 l, 8.15 l, ve 14.14 l değerlerini almıştır.

Her kümenin eleman dağılımı Şekil 5.6'daki gibi gösterilmiştir. Buna göre şehir içi yakıt tüketiminde en kalabalık kategoriler, k-ortalamalar ile cluster-2, two-step ile cluster-3 bulunmuştur. Şekil 5.6a'da k-ortalamalar belirlenen kümelerin yüzdeliği şu sıraya göre azalmaktadır: cluster-3 cluster-4, cluster-2, cluster-1. Aynı şekilde, iki aşamalı kümeleme algoritmasıyla belirlenen kümelerin sırası Şekil 5.6b'deki gibi cluster-2, cluster-5, cluster-3, cluster-4, cluster-1 olmuştur.

5.3.2. Şehir dışı yakıt tüketimi

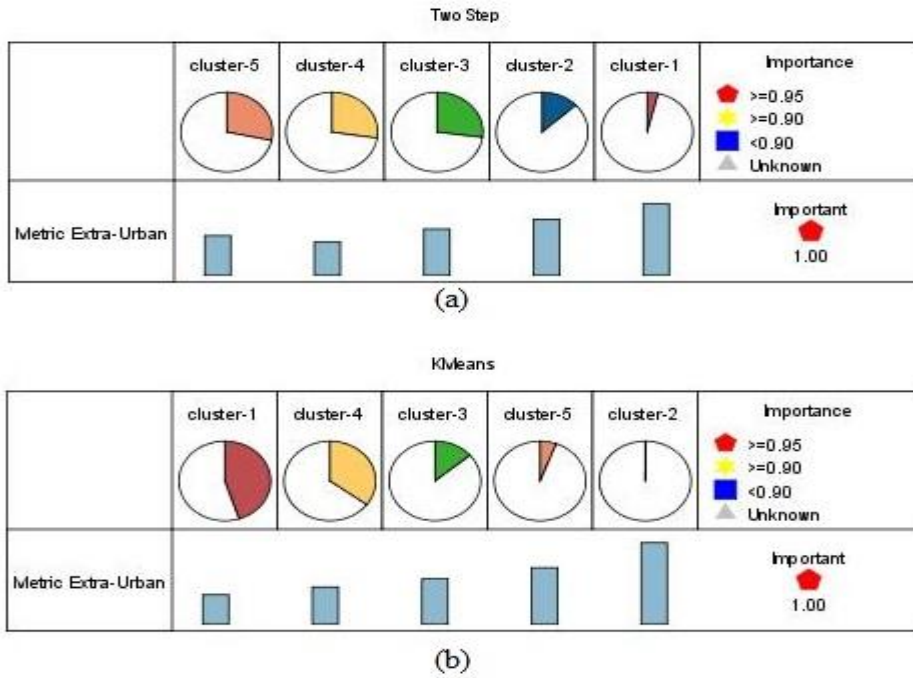
Şehir dışı yakıt tüketim çevrimi (extra urban cycle) araçların şehir içi ortamın dışındaki yollarda kullanımını göstermeye yöneliktir (Offices ve diğerleri, 2016). Hemen şehir içi çevrimin ardından gerçekleştirilir ve kabaca hızlanmalardan, yavaşlamalardan ve bazı boşalmaların dışında kalan yarı sabit hızla sürüşü de içerir. Maksimum hız saatte 120 kilometre, ortalama hız saatte 63 kilometre ve toplam mesafe 7 kilometredir.

Kümeleme algoritmaları verideki şehir dışı yakıt tüketim değerlerine (metric extra urban cold) uygulandığında Şekil 5.7'deki sonuçlar elde edilmiştir. Her iki yöntemde de 5 kategori bulunmuştur fakat bu grupların merkez değerleri farklıdır. k-ortalamalar ile bulunan cluster-2'de sadece 2 örnek bulunmasına rağmen analiz sonuçlarını büyük ölçüde değiştirdiği için değişiklik yapılmamış ve k 5 alınmıştır. İki aşamalı kümelemede en çok örneği içeren cluster-5 (345) olurken, k-ortalamlarda cluster-1 (552) olmuştur. İki aşamalı kümeleme algoritmasında bulunan şehir dışı yakıt tüketim gruplarının merkezleri sırasıyla 8.06 l, 6.32 l, 5.26 l, 3.79 l ve 4.48 l olmuştur. Bu değerler k-ortalamlarda sırasıyla 4.01 l, 11.05 l, 6.16 l, 5.06 l, ve 7.68 litredir.



Şekil 5.7. Şehir dışı yakıt tüketimi için (a) iki aşamalı kümeleme algoritması ve (b) k-ortalamalar algoritması ile belirlenen kümeler

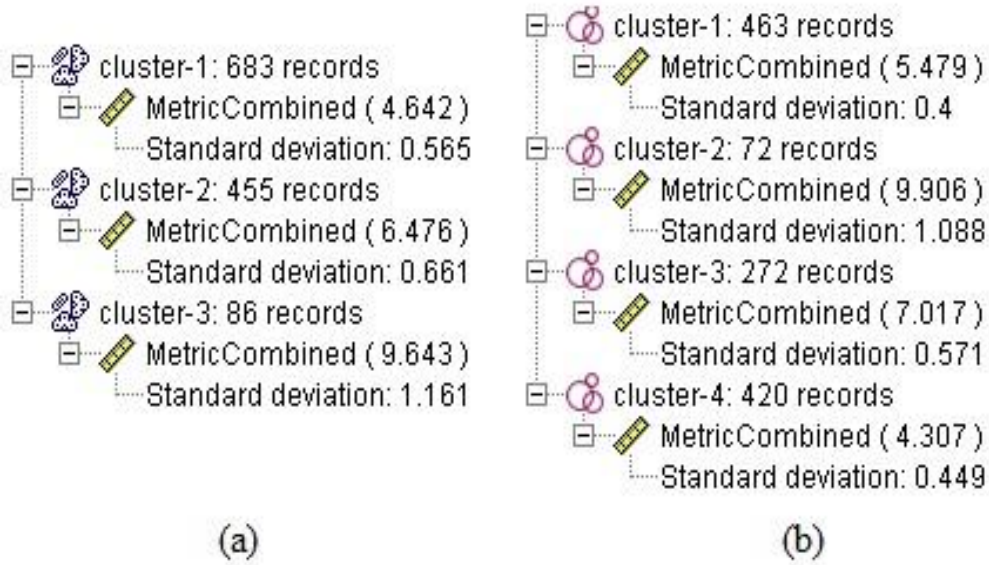
Şehir dışı yakıt tüketiminin kategorizasyonunda her kümenin eleman dağılımı Şekil 5.8’de gösterilmiştir. İki aşamalı kümeleme algoritmasıyla bulunan en kalabalık küme Şekil 5.8a’da görüldüğü gibi, cluster-1 olmuştur ve sırasıyla cluster-2, cluster-3, cluster-5, cluster-4 takip etmektedir. Aynı şekilde, k-ortalamalar algoritmasıyla belirlenen kümelerin eleman sayısı cluster-2, cluster-5, cluster-3, cluster-4, ve cluster-1 sırasında azalmaktadır (Şekil 5.8b).



Şekil 5.8. Şehir dışı yakıt tüketimi için (a) iki aşamalı kümeleme algoritması ve (b) k-ortalamalar algoritması ile bulunan kümelerin dağılımlarının karşılaştırılması

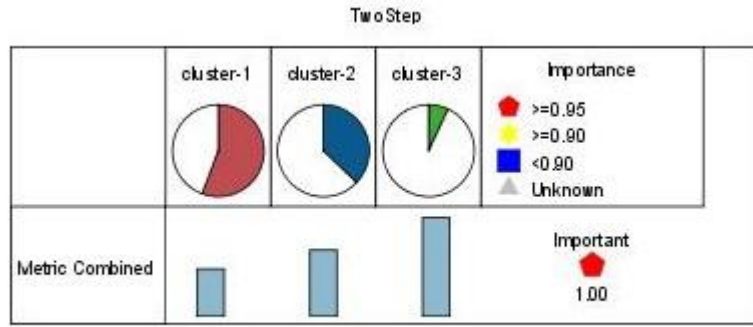
5.3.3. Kombine yakıt tüketimi

Her 100 kilometredeki litre cinsinden yakıt tüketimi, şehir içi ve şehir dışı çevrimler bir arada dikkate alınarak hesaplanmaktadır. Verideki kombine çevrim değerleri de, mesafelerine göre ağırlıklandırılmış şehir içi çevrim ve şehir dışı çevrim değerlerinin ortalamaları alınarak hesaplanmıştır.

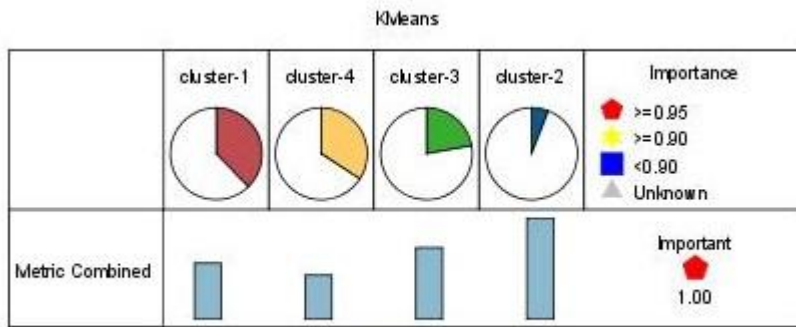


Şekil 5.9. Kombine yakıt tüketimi için (a) iki aşamalı kümeleme algoritması ve (b) k-ortalamalar algoritması ile belirlenen kümeler

Kombine yakıt tüketimi için iki aşamalı kümeleme algoritmasıyla 3, k-ortalamalar algoritmasıyla 4 grup elde edilmiştir (Şekil 5.9). Her iki algoritmayla bulunan cluster-1, en çok örnek içeren kümeler olmuştur. İki aşamalı kümeleme algoritmasıyla bulunan kümelerin merkezleri sırasıyla 4.64 l, 6.48 l ve 9.64 l olmuştur (Şekil 5.9a). k-ortalamalarda ise 5.48 l, 9.91 l, 7.02 l, ve 4.31 litredir (Şekil 5.9b). Bu kümelerin dağılımları Şekil 5.10'da gösterilmiştir. Buna göre; iki aşamalı kümeleme algoritmasıyla bulunan kümelerin eleman sayıları cluster-3, cluster-2, cluster-1 şeklinde artmaktadır (Şekil 5.10a). k-ortalamalar algoritmasında bu sıra cluster-2, cluster-3, cluster-1, cluster-4 olmuştur (Şekil 5.10b).



(a)



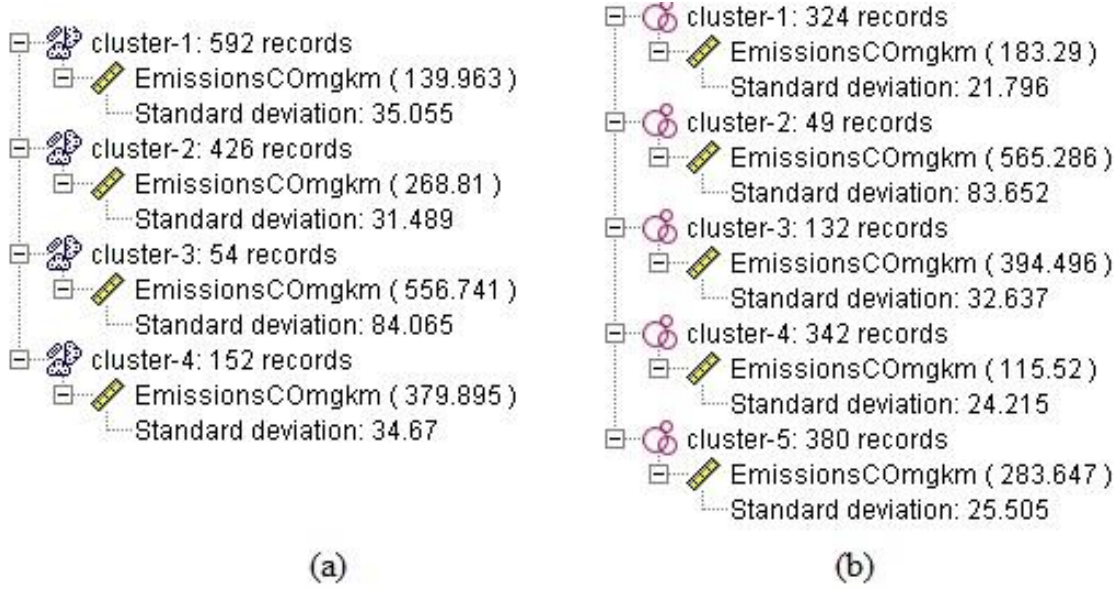
(b)

Şekil 5.10. Kombine yakıt tüketimi için (a) iki aşamalı kümeleme algoritması ve (b) k-ortalamalar algoritması ile bulunan kümelerin dağılımlarının karşılaştırılması

5.3.4. Karbon monoksit (CO) salınımı

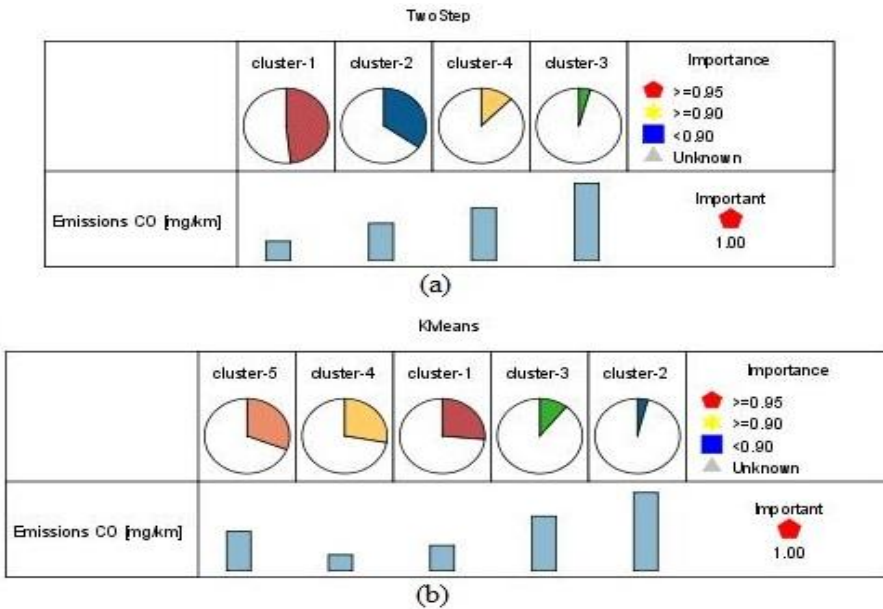
Karbon monoksit, kandaki oksijenin organlara taşınma kapasitesini azaltır. Bu sebeple, otomobiller tarafından üretilen karbon monoksit (CO) miktarı her 100 kilometrede gram cinsinden değerlendirilmiştir (Band, 2016).

Verideki CO salınım miktarları iki aşamalı kümeleme algoritması ile 4, k-ortalamalar algoritmasıyla 5 kategoriye ayrılmıştır (Şekil 5.11). En çok örnek iki aşamalı algoritmada cluster-1 (140 g)'e ait olmuş, diğer 3 kümenin merkezleri sırasıyla 270 g, 556 g, 380 g olmuştur (Şekil 5.11a). k-ortalamalarda ise merkez değeri 284 g olan cluster-5 380 örnekle en kalabalık grup olmuştur (Şekil 5.11b). Diğer kümeler sırasıyla 183 g, 565 g, 395 g, 115 g, 284 g değerlerini almaktadır.



Şekil 5.11. Karbon monoksit (CO) salınımı için (a) iki aşamalı kümeleme algoritması (b) k-ortalamlar algoritması ile belirlenen kümeler

Şekil 5.12’de, bulunan kategorilerin dağılımları belirtilmektedir. İki aşamalı kümeleme algoritmasıyla bulunan kümelerin eleman sayıları cluster-3, cluster-4, cluster-2, cluster-1 şeklinde artmaktadır (Şekil 5.12a). k-ortalamlarda ise bu sıra cluster-2, cluster-3, cluster-1, cluster-4, cluster-5 şeklindedir (Şekil 5.12b).

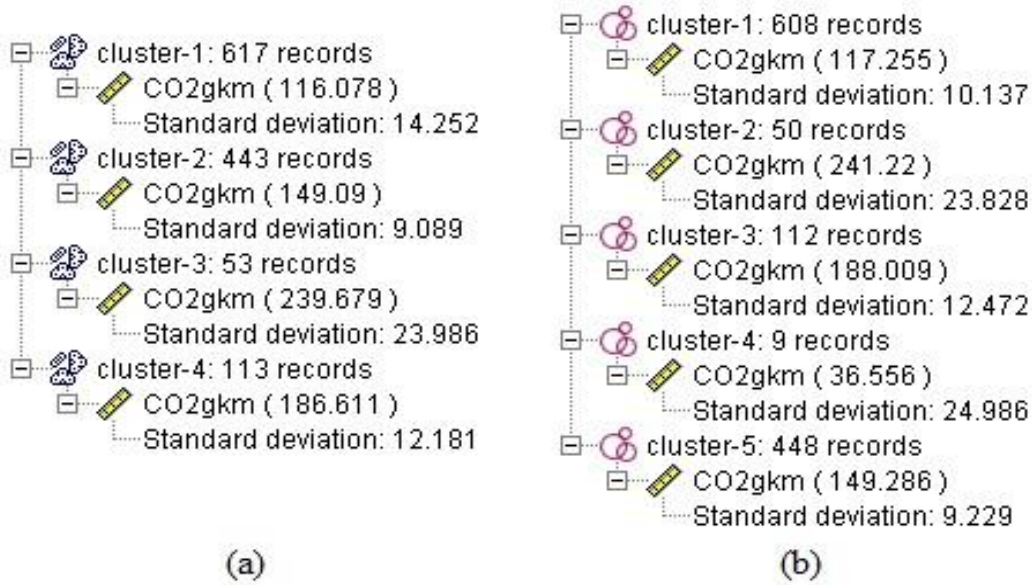


Şekil 5.12. Karbon monoksit (CO) salınımı için (a) iki aşamalı kümeleme algoritması ve (b) k-ortalamlar algoritması ile bulunan kümelerin dağılımlarının karşılaştırılması

5.3.5. Karbon dioksit (CO₂) salınımı

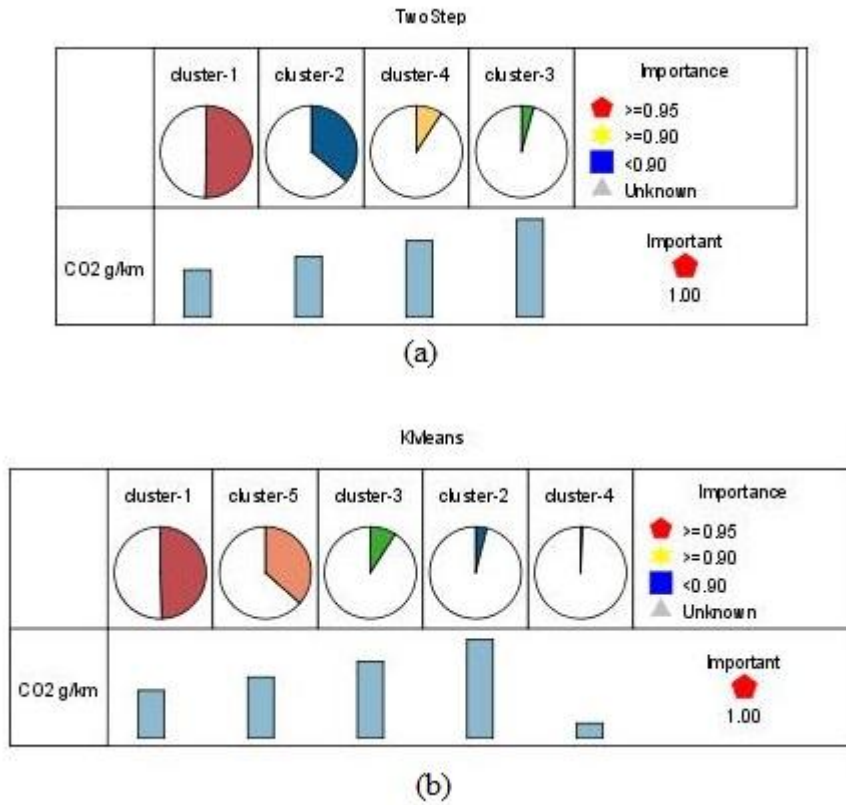
Petrol ve dizel, araçlarda enerji için kullanıldığında temel yan ürünler su buharı ve karbon dioksit (CO₂) olmaktadır. İklim değişikliklerine katkıda bulunan en önemli sera gazı da CO₂'tir. 6000 millik mesafede bir otomobilin ürettiği CO₂ miktarı neredeyse kendi ağırlığına eşittir. 1 Mart 2001'den sonra kaydedilen otomobiller CO₂ salınım miktarına ve kullanılan yakıt tipine bakılarak vergilendirilmektedir (Band, 2016).

Karbondioksit (CO₂) salınımı için belirlenen kategoriler Şekil 5.13'teki gibi iki aşamalı kümeleme algoritmasıyla 4, k-ortalamalar algoritmasıyla 5 grupta toplanmıştır. En çok örnek her iki yöntemde de cluster-1'e ait olmuştur. İki aşamalı kümeleme algoritmasında bulunan kategorilerin merkezleri 116 g, 149 g, 240 g, 187 g, k-ortalamalarda ise 117 g, 241 g, 188 g, 37 g, 150 g olmuştur.



Şekil 5.13. Karbon dioksit (CO₂) salınımı için (a) iki aşamalı kümeleme algoritması ve (b) k-ortalamalar algoritması ile belirlenen kümeler

Bulunan kümelerin eleman sayıları Şekil 5.14'teki gibi iki aşamalı kümeleme algoritması için cluster-3, cluster-4, cluster-2, cluster-1 ve k-ortalamalar için cluster-2, cluster-3, cluster-5, cluster-1, cluster-4 olmuştur.

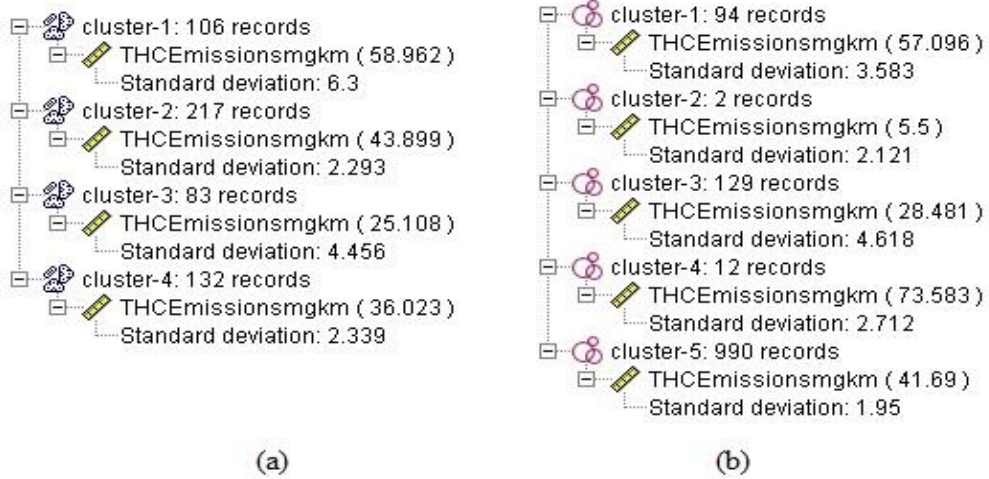


Şekil 5.14. Karbon dioksit (CO₂) salınımı için (a) iki aşamalı kümeleme algoritması ve (b) k-ortalamalar algoritması ile bulunan kümelerin dağılımlarının karşılaştırılması

5.3.6. Hidrokarbon (THC) salınımı

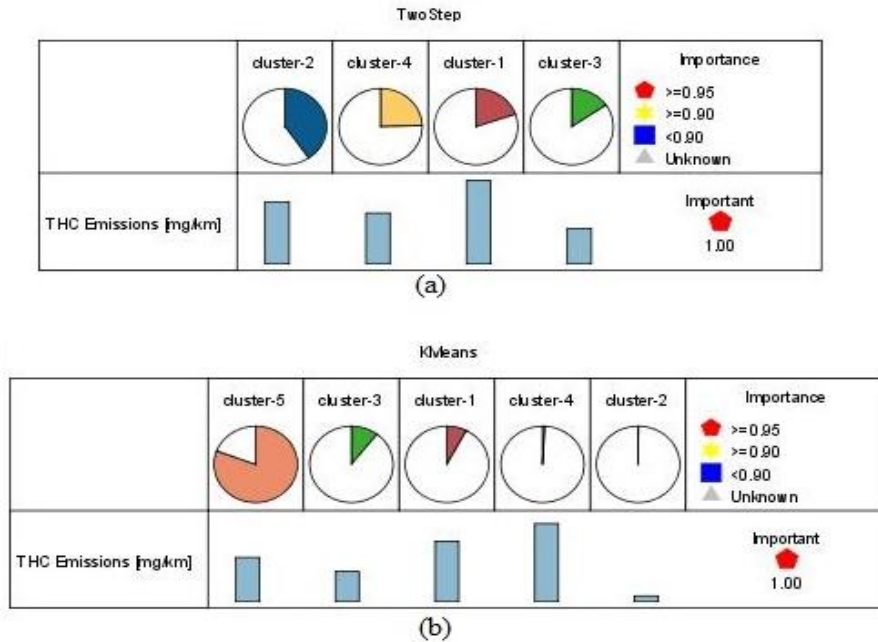
Hidrokarbonlar ozon tabakasının şekillendirilmesine katkıda bulunmaktadır. Bazı çeşit hidrokarbonlar indirekt sera gazları olabildiği gibi, kanserojen özellik de gösterebilmektedir (Band, 2016). Toplam hidrokarbon emisyonu (THC) kilometre başına gram ölçümüyle hesaplanmaktadır.

Veride yer alan THC emisyon değerleri iki aşamalı kümeleme algoritmasıyla 4, k-ortalamalar ile 5 kategoriye ayrılmıştır (Şekil 5.15). İki aşamalı algoritmada en çok örnek cluster-2'ye dahildir ve bulunan kategorilerin merkezleri sırasıyla 59 g, 44 g, 25 g, 36 g şeklinde değişmektedir (Şekil 5.15a). Şekil 4.13b'de k-ortalamalar ile bulunan kümelerde ise hidrokarbon kategorileri 57 g, 5.5 g, 28.5 g, 73.5 g ve 42 g merkezlerinde toplanmıştır.



Şekil 5.15. Hidrokarbon (THC) emisyonu için (a) iki aşamalı kümeleme algoritması (b) k-ortalamalar algoritması ile belirlenen kümeler

THC emisyonu için bulunan bu kümelerin dağılımları Şekil 5.16'da gösterilmiştir. Buna göre iki aşamalı kümeleme algoritmasıyla bulunan 4 kategorinin eleman sayılarına göre sırası cluster-1, cluster-2, cluster-4, cluster-3'tür (Şekil 5.16a). Bu sıra, k-ortalamalar algoritmasıyla cluster-4, cluster-1, cluster-5, cluster-3, cluster-2 şeklinde bulunmuştur.

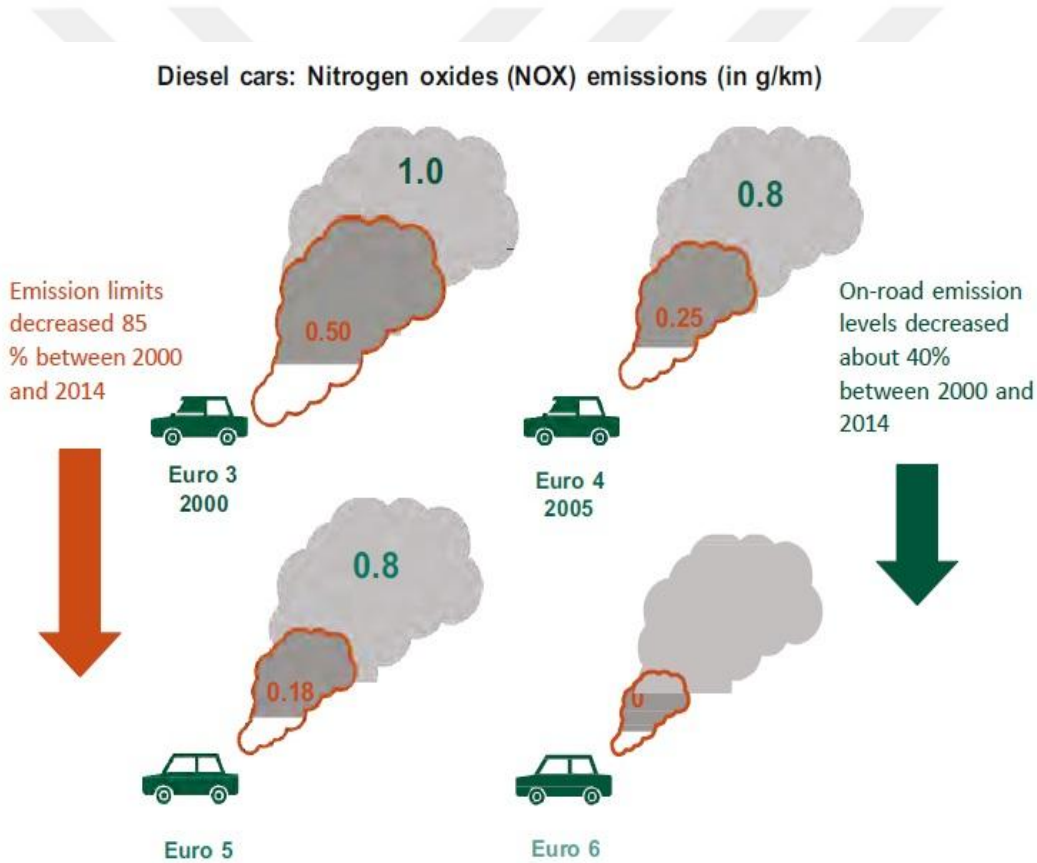


Şekil 5.16. Hidrokarbon (THC) emisyonu için (a) iki aşamalı kümeleme algoritması ve (b) k-ortalamalar algoritması ile bulunan kümelerin dağılımlarının karşılaştırılması

5.3.7. Nitrojen oksit (NO_x) emisyonu

Nitrojen oksitler (NO_x) sađlık üzerinde ciddi olumsuz etkilere neden olabilir. Özellikle, nitrojen dioksit (NO_2) prematüre ölümlere kadar giden önemli solunum problemlerine sebep olmaktadır (Franco ve diğlerleri, 2014).

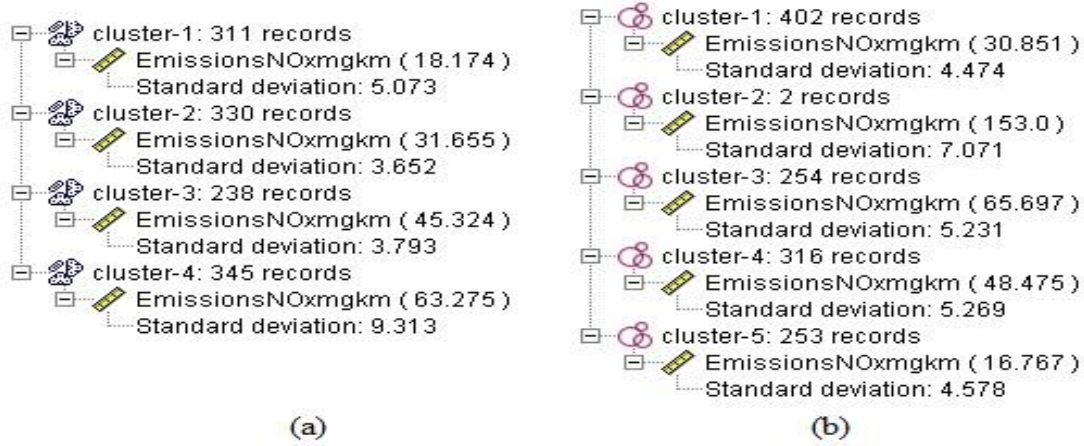
2014'te çıkan mevcut Euro 6 standardında nitrojen oksit (NO_x) limiti yaklaşık olarak %90 oranında düşürülmüştür. Ancak, son kayıtlara bakıldığında hava kalitesinin geliştirilememesindeki temel sebep, araçların gerçek hayatta test ortamında ölçülen değlerleri göstermemesidir (The UK Department for Transport, 2016) (Şekil 5.17).



Şekil 5.17. Standartlardaki emisyon limitleri ve yolda ölçülen değlerler arasındaki fark (Franco ve diğlerleri, 2014)

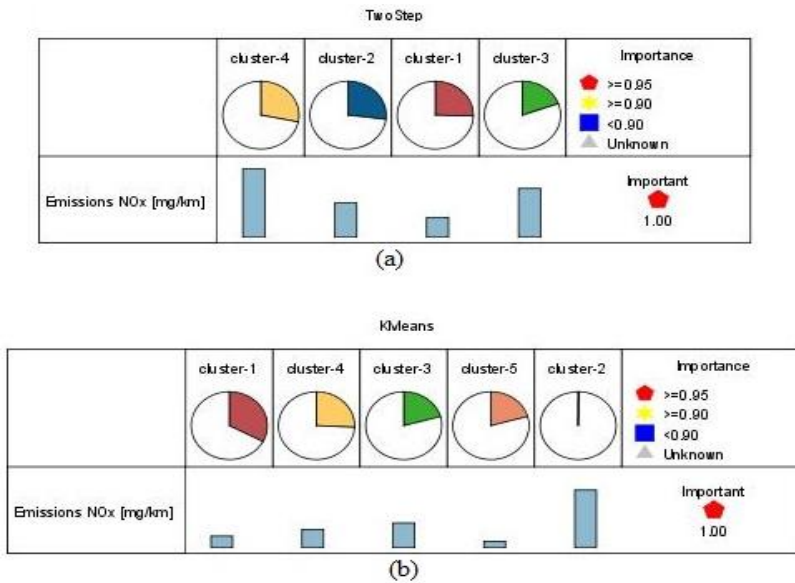
Veri kümesinde bulunan nitrojen oksit değlerleri iki aşamalı kümeleme algoritmasıyla 4, k-ortalamalar algoritmasıyla 5 kategoriye ayrılmıştır (Şekil 5.18). Şekil 5.18a'da görüldüğü gibi, iki aşamalı kümeleme algoritmasıyla bulunan küme merkezleri sırasıyla 18 g, 32 g, 45 g, 63 g olmuştur ve en çok örnek cluster-4'te yer almaktadır. Aynı şekilde k-ortalamalar ile

bulunan küme merkezleri 31 g, 153 g, 66 g, 48.5 g ve 17 gramdır ve en çok örnek Şekil 5.18b'deki gibi cluster-1'de yer almaktadır.



Şekil 5.18. Nitrojen Oksit (NO_x) emisyonu için (a) iki aşamalı kümeleme algoritması (b) k-ortalamalar algoritması ile belirlenen kümeler

Nitrojen oksit (NO_x) emisyonu için bulunan her kümenin dağılımı Şekil 5.19'da gösterilmiştir. İki aşamalı kümeleme algoritmasıyla bulunan kümelerin eleman sayısına göre azalan sırası cluster-4, cluster-2, cluster-1, cluster-3 olmuştur (Şekil 5.19a). k-ortalamalar metodu ise kümeler eleman sayısına göre Şekil 5.19b'deki gibi cluster-2, cluster-5, cluster-3, cluster-4, cluster-1 şeklinde sıralamıştır.

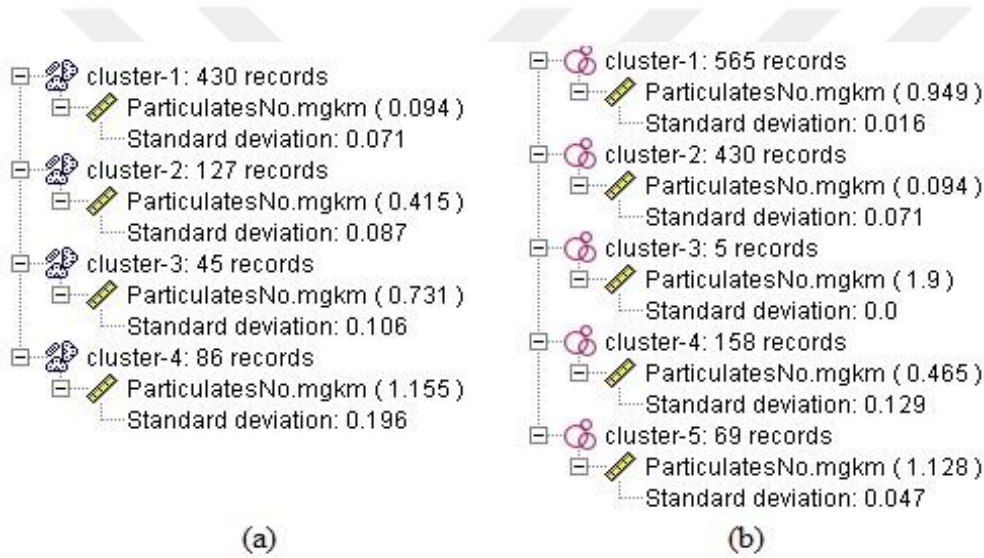


Şekil 5.19. Nitrojen oksit (NO_x) emisyonu için (a) iki aşamalı kümeleme algoritması ve (b) k-ortalamalar algoritması ile bulunan kümelerin dağılımlarının karşılaştırılması

5.3.8. Paracık emisyonu

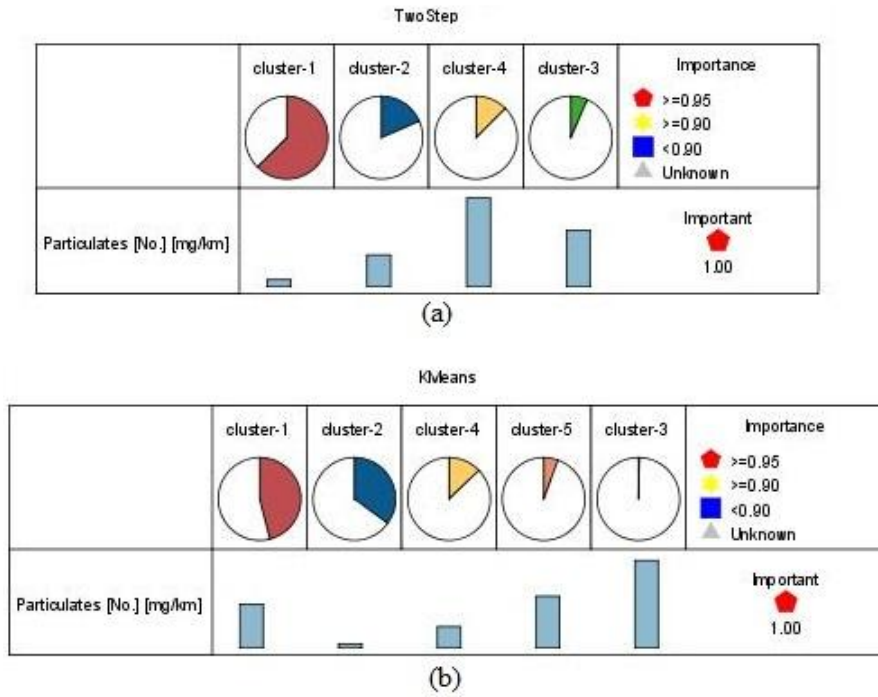
İnce paracıklar (özellikle 10 mikrondan küçük olanlar) akciğere girerek çeşitli solunum ve kalp-damar rahatsızlıklarına yol açmaktadır (Band, 2016). Bu yüzden otomobillerde her 100 kilometredeki paracık emisyonu ölçülmektedir.

Veri kümesinde bulunan paracık salınım değeri iki aşamalı kümeleme algoritmasıyla 4, k-ortalamlar ile 5 kategoriye ayrılmıştır (Şekil 5.20). İlk yöntemde bulunan küme merkezleri 0.09 g, 0.41 g, 0.73 g ve 1.15 g olurken, k-ortalamlar ile 0.95 g, 0.09 g, 1.9 g, 0.46 g ve 1.13 g bulunmuştur. Her iki yöntemde de örneklerin çoğu cluster-1'e dahildir.



Şekil 5.20. Paracık emisyonu için (a) iki aşamalı kümeleme algoritması (b) k-ortalamlar algoritması ile belirlenen kümeler

Bulunan kümelerdeki dağılımlar Şekil 5.21'de gösterilmiştir. İçerdiği örnek sayısının 5 olmasına rağmen k-ortalamlar yöntemiyle bulunan cluster-3 analizde etkili bulunmasından dolayı analizde yer almıştır.

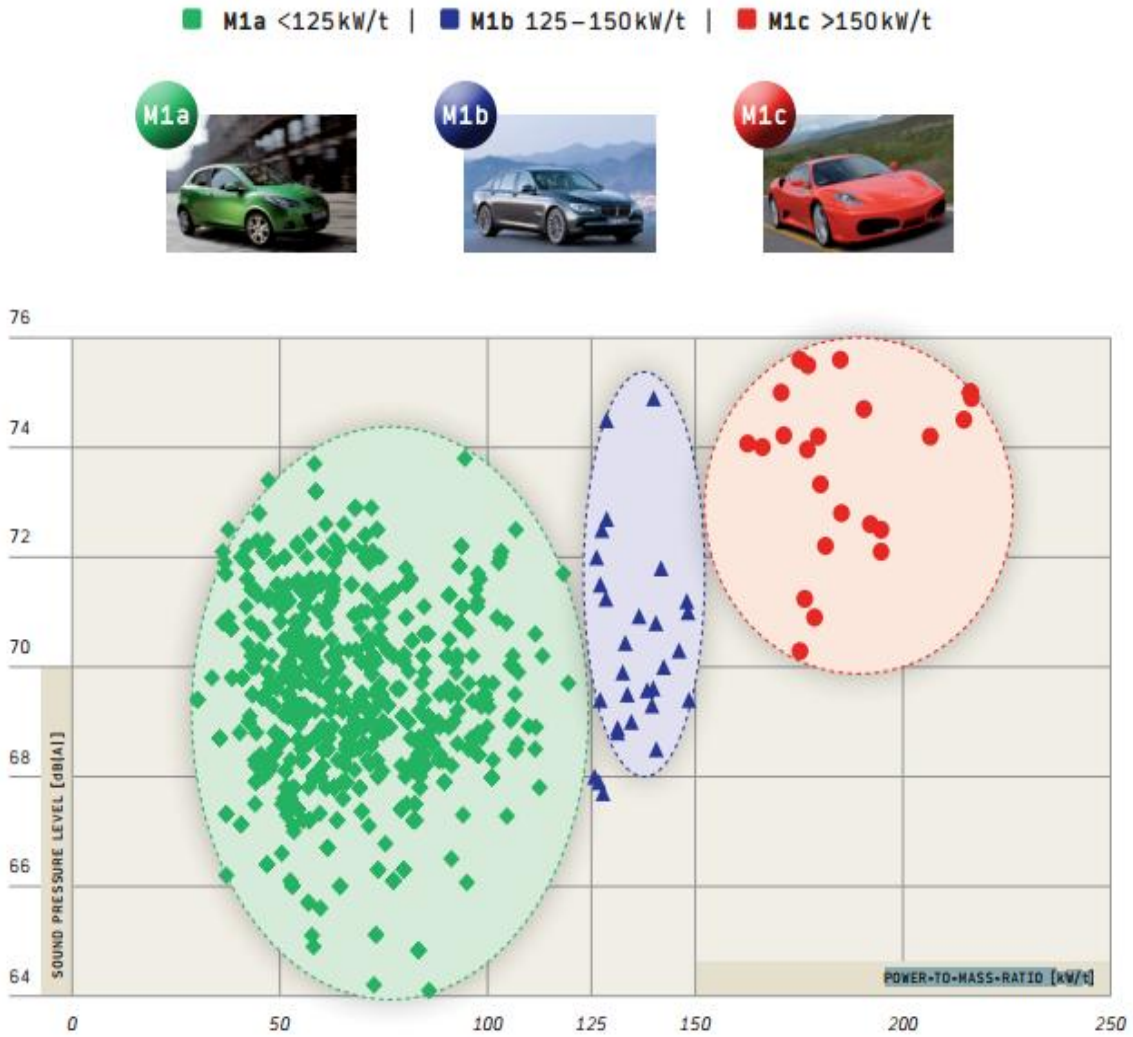


Şekil 5.21.Parçacık emisyonu için (a) iki aşamalı kümeleme algoritması ve (b) k-ortalamalar algoritması ile bulunan kümelerin dağılımlarının karşılaştırılması

5.3.9. Gürültü seviyesi

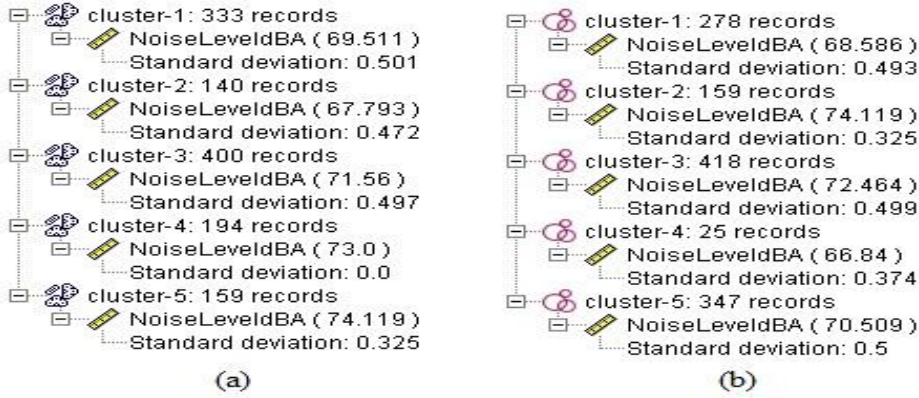
Küçük hatchbacklerden spor araçlara, geniş ve lüks sedanlardan 9 koltuklulara kadar bütün özel otomobiller mevzuatta M1 kategorisine girmektedir. Bu kategoride de ses seviyesi ile ilgili çeşitli gereklilikler mevcuttur (ACEA, 2012). Mevcut durumu geliştirmek için çeşitli araç ailelerinin güç/kütle oranına dayanan özel bir otomobil sınıflandırma sistemi önerilmiştir. Bu sınıflandırma var olan M1 kategorisini 3 segmente bölmekte ve araçların %99'u yeşil 'normal' grubunda yer almaktadır. Şekil 5.22, M1 kategorisinin 3 tipi için araçların ses basıncı aralıklarını göstermektedir.

NEW CATEGORISATION PROPOSED BY ACEA



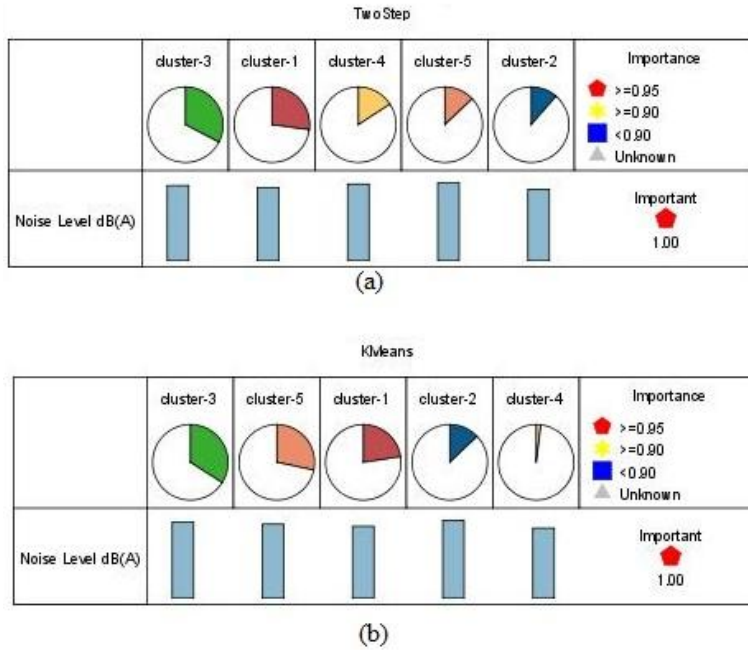
Şekil 5.22. M1 tip özel araçlarının ses seviyesi kategorileri (ACEA, 2012)

Özel otomobillerin mevcut gürültü test standardı, AB 92/97 direktiflerinde detaylı olarak açıklandığı gibi, aracın test alanında 50 km/saat hızla sürülmesi ve tam gazla hızlanmasını içerir. Ulaşılan maksimum ses seviyesi ayarlanan mesafede bir mikrofona ölçülür. Gürültü seviyeleri gürültü ölçerin A ölçeğinde desibel cinsinden gösterilir. A ölçeği, ses değerinin okumada ağırlıklı olarak tasarlandığından insan kulağı ile algılanan değere çok yakındır (Band, 2016).



Şekil 5.23. Gürültü seviyesi için (a) iki aşamalı kümeleme algoritması (b) k-ortalamalar algoritması ile belirlenen kümeler

Mevcut gürültü seviyeleri verisi, her iki kümeleme algoritmasıyla da 5 kategoriye ayrılmıştır (Şekil 5.23). İki aşamalı kümeleme algoritmasıyla belirlenen kümelerden en kalabalık olanı 400 örnekle cluster-3 olmuştur. Küme merkezleri sırasıyla 69.5 dB[A], 67.8 dB[A], 71.6 dB[A], 73 dB[A] ve 74. dB[A] olmuştur. K-ortalamalar ile bulunan kümelerden en çok örneğe sahip olanı aynı şekilde cluster-3'tür. Elde edilen kümelerin merkezi 68.6 dB[A], 74.1 dB[A], 72.5 dB[A], 66.8 dB[A], and 70.5 dB[A] olmuştur. Gürültü seviyesi için belirlenen kümelerin dağılımı Şekil 5.24'te gösterilmiştir.



Şekil 5.24. Gürültü seviyesi için (a) iki aşamalı kümeleme algoritması ve (b) k-ortalamalar algoritması ile bulunan kümelerin dağılımlarının karşılaştırılması

En iyi kümeleme algoritmasının seçimi

Kümeleme aşamasında, k-ortalamlar algoritmasındaki kümelerin çıktıları ve kalitesi doğru sayıda kümenin seçimine bağlıdır. Optimal küme sayısı da Davies-Bouldin Index, Silhouette genişliği veya Dunn Index gibi göstergelere bakılarak belirlenir.

Dunn Index, sıklığı (kümelerin veri noktaları arasındaki maksimum mesafe) ve ayrılmayı (kümeler arasındaki minimum mesafe) ölçer. Bu ölçüm yardımıyla veri kümesinde doğru sayıda küme elde edilir ve indexin maksimum değeri doğru ayırmanın yapıldığını gösterir (Pakhira, Bandyopadhyay ve Maulik, 2004).

Bu tezde kullanılan Dunn Index Matlab kodları aşağıdaki gibidir:

```
function DI=dunns(clusters_number, distM,ind)
%%Dunn's index for clustering compactness and separation measurement
% dunns(clusters_number, distM,ind)
% clusters_number = Number of clusters
% distM = Dissimilarity matrix
% ind = Indexes for each data point aka cluster to which each data point
% belongs
i=clusters_number;
denominator= [];
for i2= 1:i
    indi=find(ind==i2);
    indj=find(ind~=i2);
    x=indi;
    y=indj;
    temp=distM(x, y);
    denominator=[denominator;temp(:)];
end
num=min(min(denominator));
neg_obs=zeros(size(distM,1), size(distM,2));
for ix= 1:i
    idxs=find(ind==ix);
```

```

    neg_obs(indxs, indxs)=1;
end
dem=neg_obs.*distM;
dem=max(max(dem));
DI=num/dem;
end

```

Bu göstergelerin kullanılmasının sebebi; bir kümedeki elemanların birbirine benzer, diğer kümelerdekilerden farklı olması için daha anlamlı kümeler elde etmektir. Bu yüzden optimal küme sayıları, Dunn Indexin k-ortalamlar algoritmasında kullanılmasıyla belirlenmiştir.

5.4. Algoritma Performanslarının Değerlendirilme Aşaması

Çalışmada incelenen üç farklı sınıflandırma yönteminin (sinir ağları, bayes ağları ve C5.0) yakıt tüketimi, emisyon ve gürültü seviyesine göre performansının değerlendirilmesinde Clementine’da Evaluation Node kullanılmıştır. Evaluation Node sayesinde öngörü modelleri değerlendirilerek karşılaştırılabilmekte ve uygulama için en iyisi seçilebilmektedir. Değerlendirme grafikleri, olası çıktıların tahmininde modellerin nasıl bir yol izlediğini göstermektedir. Bunlar, tahmin değerleri ve tahmin güvenilirliğine dayanarak kayıtları sıralar, kayıtları eşit büyüklükteki gruplara (quantiles) ayırır. Daha sonra her niceliğe göre iş kriterlerinin değerini en yüksekte en düşüğe doğru çizer. Farklı modeller aynı grafikteki bağımsız çizgilerle gösterilir.

Çıktılar elde edilen belirli değerin veya değer aralığının isabetli olarak tanımlanmasıyla işlenir. İsabetli değerler genellikle bazı alanlarda başarıyı (müşteriye satış gibi) veya ilgilenilen bir olayın başarısını (örneğin; tıpta bir tanı) belirtir. İsabet kriteri diyalog kutusunun Seçenekler kısmında tanımlanabilir veya varsayılan kriter seçilerek kullanılabilir.

Her biri farklı değerlendirme kriterlerini belirten 5 tipte değerlendirme grafiği bulunmaktadır.

Kazanç Grafiği (Gain Chart): Kazançlar, her nicelikte ortaya çıkan toplam isabet oranıdır. Nicelikteki isabet sayısının toplam isabet sayısına bölünüp 100 ile çarpılmasıyla hesaplanır.

Yükseltme Grafiği (Lift Chart): İsaletli olan kayıtların oranıyla öğrenme verisindeki toplam isabet oranını karşılaştırır. Nicelikteki isabet sayısının, nicelikteki tüm kayıtların sayısına bölünüp 100 ile çarpılmasıyla bulunur.

Tepki Grafiği (Response Chart): Basitçe, nicelikteki isabetli olan kayıtların oranını ifade eder. Nicelikteki isabet sayısının nicelikteki kayıt sayısına bölünüp 100 ile çarpılması ile elde edilir.

Fayda Grafiği (Profit Chart): Fayda (ya da kâr), her kaydın getirisinden maliyetinin çıkartılması ile bulunur. Bir nicelikteki fayda basitçe nicelikteki tüm kayıtların kârları toplamıdır. Getirilerin sadece isabetlere uygulandığı, maliyetlerin tüm kayıtlara uygulandığı varsayılır. Kârlar ve maliyetler sabit olabilir veya verideki alanlarla tanımlanabilir. Nicelikteki kayıtların getiri toplamının nicelikteki kayıtların maliyet toplamına bölünmesiyle bulunur.

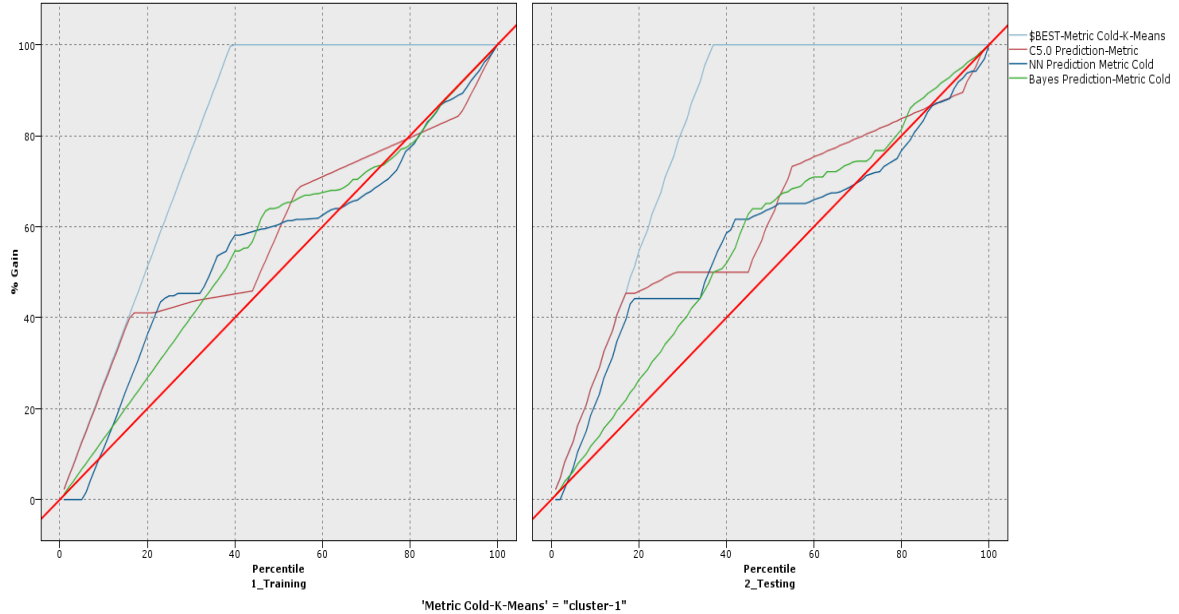
Yatırımın Getirisi Grafiği (ROI Chart): ROI (return on investment) getiri ve maliyetleri göstermesi açısından kâra benzer. ROI bir nicelik için kârı maliyetlerle karşılaştırır. Nicelikteki kârların nicelikteki maliyetlere bölünüp 100 ile çarpılmasıyla bulunur.

5.4.1. Şehir içi yakıt tüketimi

Clementine'ın ayırma aracında, verinin %80'i öğrenme için seçilmiş ve analizde test verisiyle karşılaştırılmıştır. Şehir içi yakıt tüketimi için öğrenme ve test sonuçları, öğrenmenin kullanılan segmentasyon tekniklerinin başarısını arttırdığını göstermiştir.

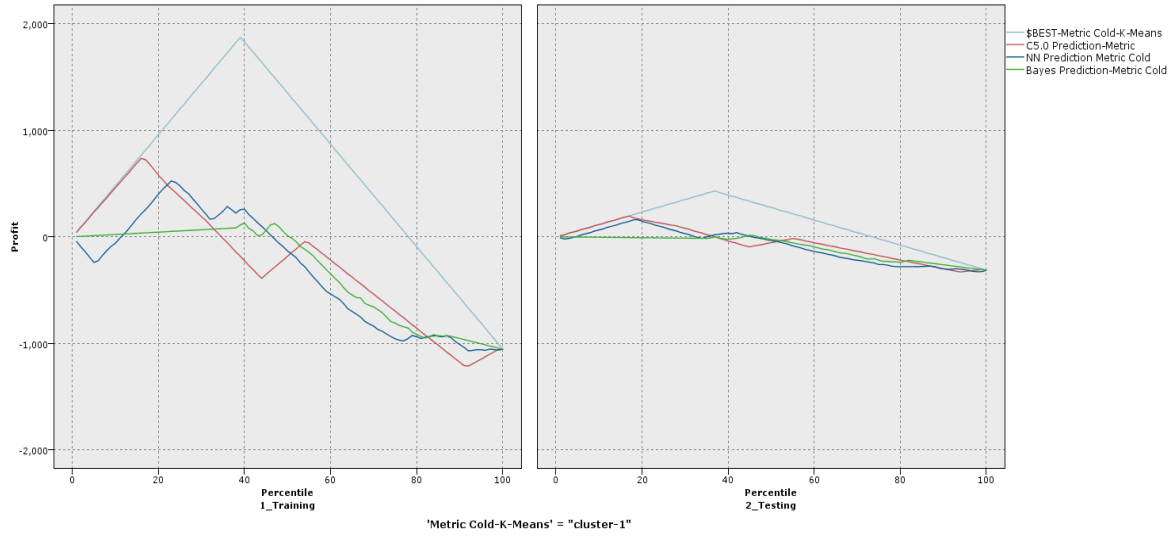
Kazanç grafiğine göre (Şekil 5.25a) tekniklerin öğrenme verisinin %40'ından itibaren %100 kazanç vermesi beklenmektedir (ideal sonuç). Tekniklerin verdiği gerçek sonuçlar ile ideal sonuçlar karşılaştırılmaktadır. Buna göre; C5.0 algoritması verinin %0-20 ve %50-80'i arasında, sinir ağları verinin %20-45'i arasında ve Bayes ağları %45-60'ı ve %80-100'ü arasında ideale en yakın sonuçları vermektedir. Öğrenme ile birlikte tüm tekniklerin kazanç değerleri yükselmiştir. Özellikle C5.0 algoritması diğer tekniklerden daha iyi sonuçlar göstermiştir.

Şekil 5.25b’de de görüldüğü gibi, öğrenme sonrasında C5.0 algoritması verinin %0-22 ve %53-78 dilimleri arasında, sinir ağı %22-45 ve %82-100 arasında ve Bayes ağı %45-53 ve %87-100 arasında en iyi sonucu vermektedir. Bu da 2260’tan fazla şehir içi yakıt tüketimi değerinin analizinde C5.0 algoritmasının daha faydalı olduğunu göstermektedir. Ayrıca, %82-87 aralığında sinir ağı ve Bayes ağı aynı tahmin sonuçlarını vermiştir. Öğrenme sonrasında, algoritmaların fayda eğrileri birbirine yaklaştığından daha homojen bir grafik elde edilmiştir. Beklenen ideal çözüm ve gerçekleşen sonuçlar da aynı şekilde birbirine yaklaşmıştır. Analizin genelinde en çok kazancı ve faydayı vermesinden dolayı (%0-35 ve %53-82) şehir içi yakıt tüketimi için kullanılması en mantıklı yöntem C5.0 algoritmasıdır.



(a)

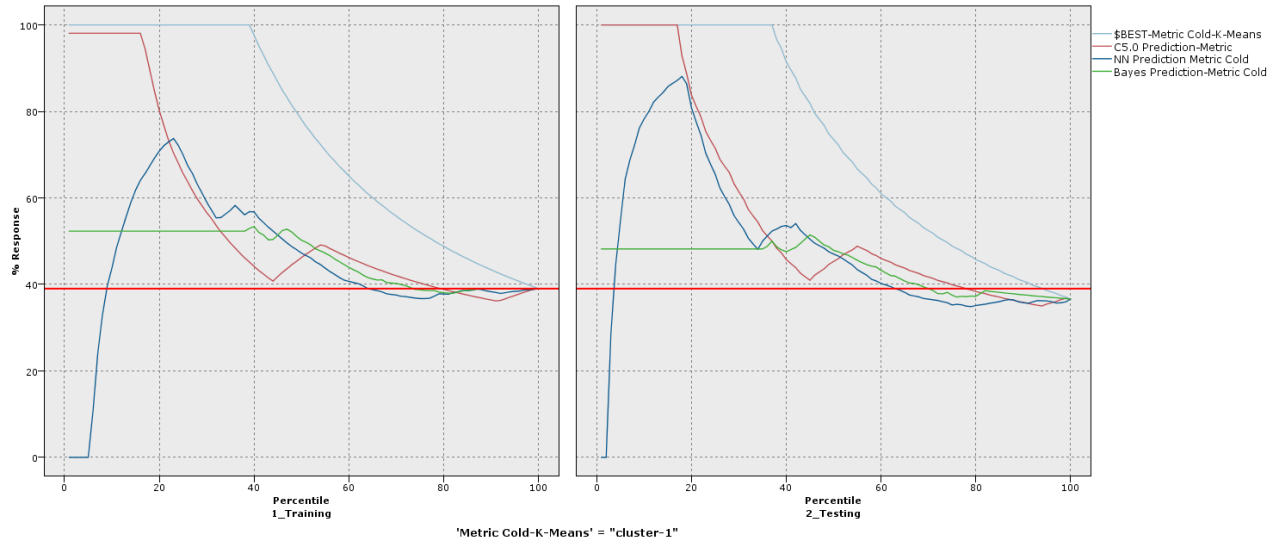
Şekil 5.25. Şehir içi yakıt tüketimi için (a) kazanç (b) kâr grafiği



(b)

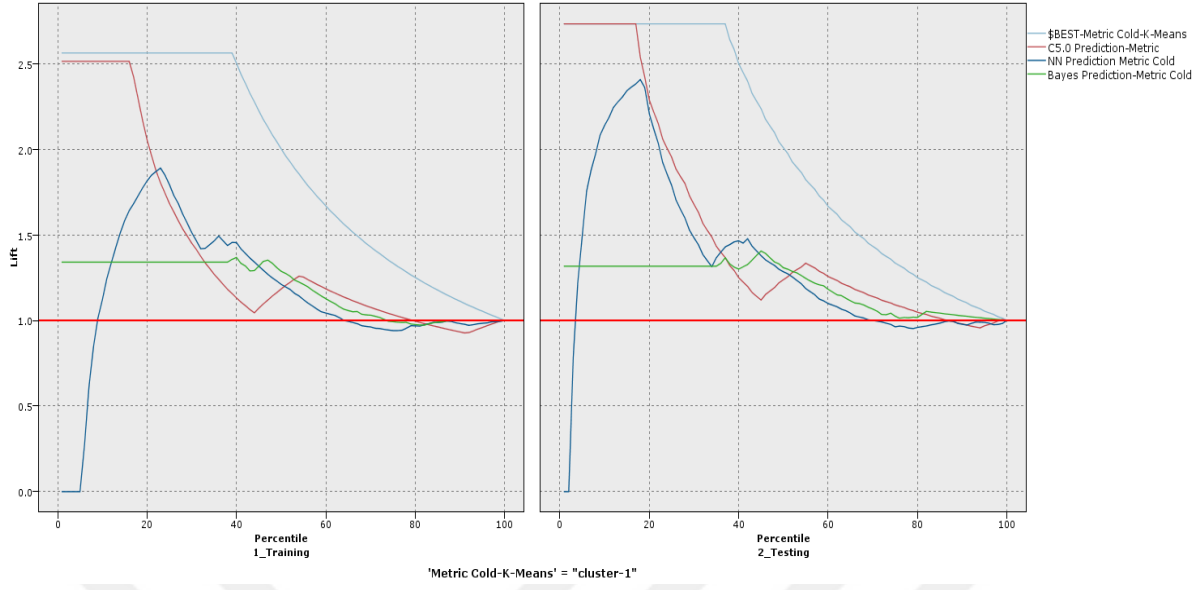
Şekil 5.25. (devam) Şehir içi yakıt tüketimi için (a) kazanç (b) kâr grafiği

Tepki, yükseltme ve ROI grafiklerinde ise neredeyse birbirinin aynısı olan sonuçlar elde edilmiştir (Şekil 5.26). Öğrenme esnasında ilk %20'lik veride ve %53-78 arasında ideal çözüme en yakın sonucu veren teknik C5.0 algoritması olmuştur. Sinir ağı %23-45 ve %83-100 arasında, Bayes ağı %45-53 ve %82'den itibaren en iyi teknik olmuştur.

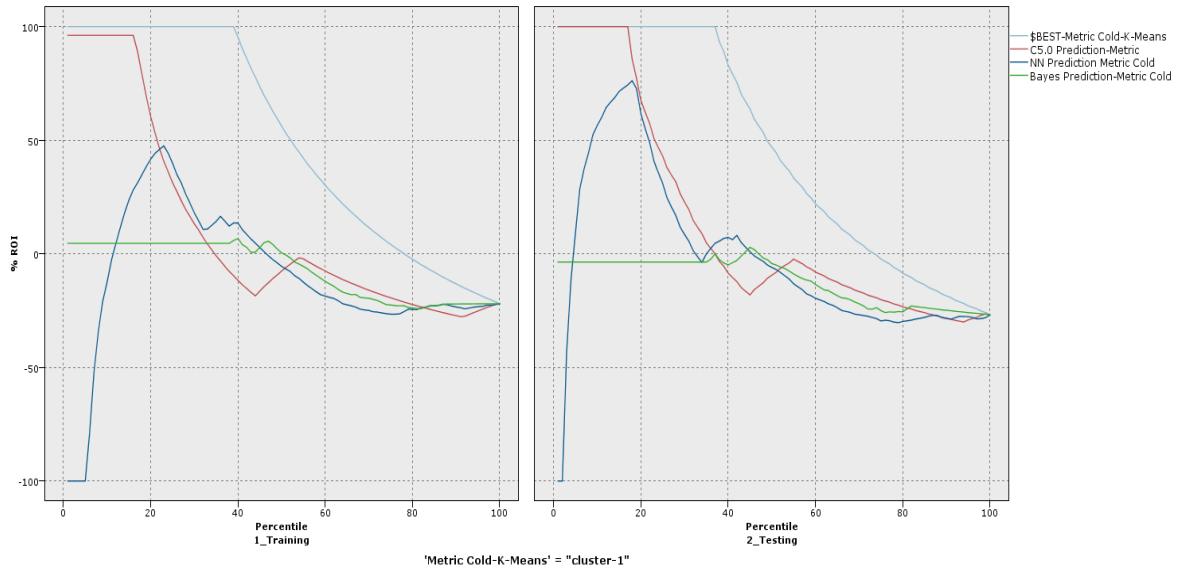


(a)

Şekil 5.26. Şehir içi yakıt tüketimi için (a) tepki, (b) yükseltme ve (c) ROI grafikleri



(b)



(c)

Şekil 5.26. (devam) Şehir içi yakıt tüketimi için (a) tepki, (b) yükseltme ve (c) ROI grafikleri

Öğrenme sonrasında, C5.0 algoritması ilk %35'lik veride ve %53-78 arasında daha iyi tepki vermektedir. Sinir ağının elde ettiği sonuçlar C5.0 algoritmasına yakın olması hatta öğrenmeden sonra %35-45 aralığında daha iyi olmasına rağmen, parametre tahmininde C5.0 algoritması daha güçlü olmuştur. Bayes ağı sadece %45-53 ve %82-100 aralığında daha iyi sonuçlar vermiş fakat aynı şekilde C5.0 algoritmasının gerisinde kalmıştır.

Genel olarak, şehir içi yakıt tüketimi için ideal çözüme en yakın sonucu veren teknik C5.0 algoritması olmuştur. Şekil 5.27’de görüldüğü gibi, C5.0 algoritması, sinir ağı ve Bayes ağının en önemli değişkenleri belirlemedeki başarıları sırasıyla %87, %82 ve %72 olmuştur.

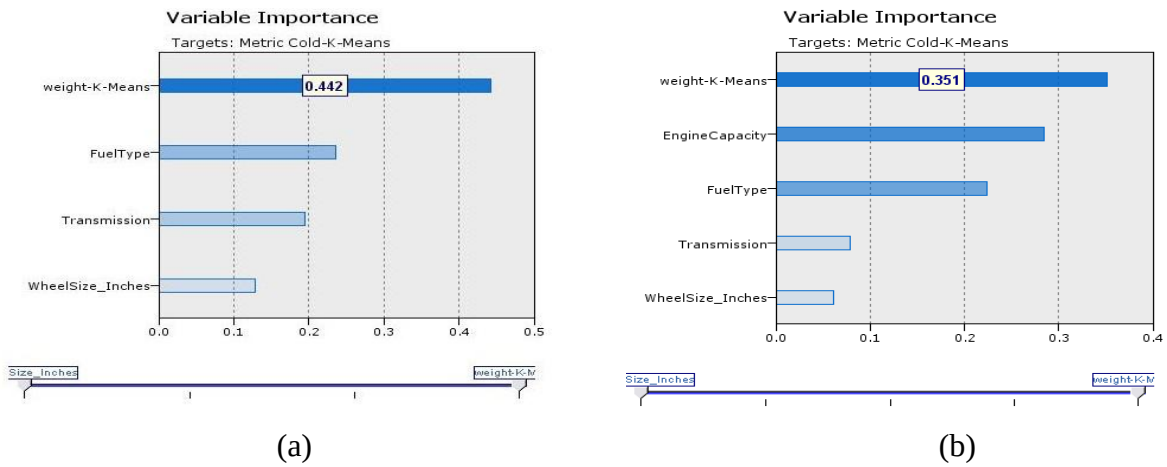
Comparing NN-Metric Cold with Metric Cold-K-Means					
'Partition'	1_Training		2_Testing		
Correct	801	83.35%	194	82.55%	
Wrong	160	16.65%	41	17.45%	
Total	961		235		

Comparing C5.0-Metric Cold with Metric Cold-K-Means					
'Partition'	1_Training		2_Testing		
Correct	835	86.89%	205	87.23%	
Wrong	126	13.11%	30	12.77%	
Total	961		235		

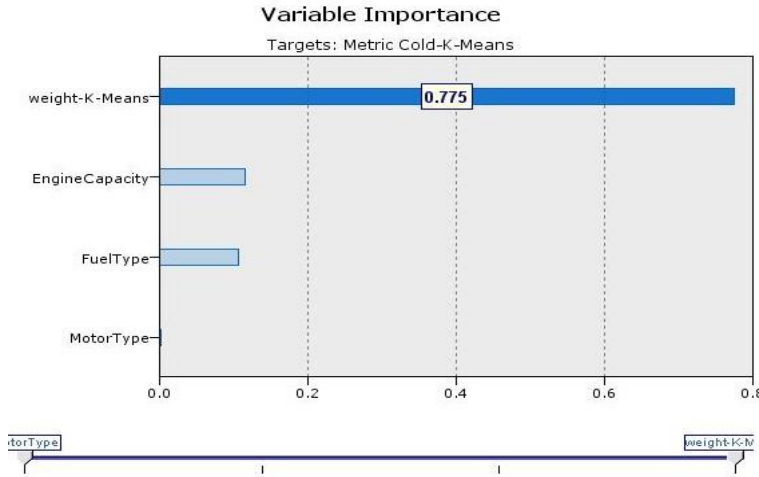
Comparing Bayes-Metric Colds with Metric Cold-K-Means					
'Partition'	1_Training		2_Testing		
Correct	748	77.84%	170	72.34%	
Wrong	213	22.16%	65	27.66%	
Total	961		235		

Şekil 5.27. Şehir içi yakıt tüketimi verilerine uygulanan tekniklerin analiz sonuçları

Tüm teknikler, şehir içi yakıt tüketimi üzerinde otomobil ağırlığının en etkili değişken olduğunu göstermektedir (Şekil 5.28). Bayes ağı ve C5.0 algoritması ikinci ve üçüncü en önemli değişkenleri sırasıyla motor kapasitesi ve yakıt tipi olarak bulmuştur. Ancak sinir ağının verdiği tahminde motor kapasitesi yer almamaktadır. Aynı şekilde C5.0 algoritması da sinir ağı ve Bayes ağının aksine vites kutusunu önemli değişkenlerden biri olarak bulmamıştır. C5.0 algoritmasıyla bulunan diğer bir etkili parametre ise motor tipi olmuştur.



Şekil 5.28. Şehir içi yakıt tüketimini etkileyen (a) sinir ağı (b) Bayes ağı (c) C5.0 algoritması ile bulunan en önemli faktörler

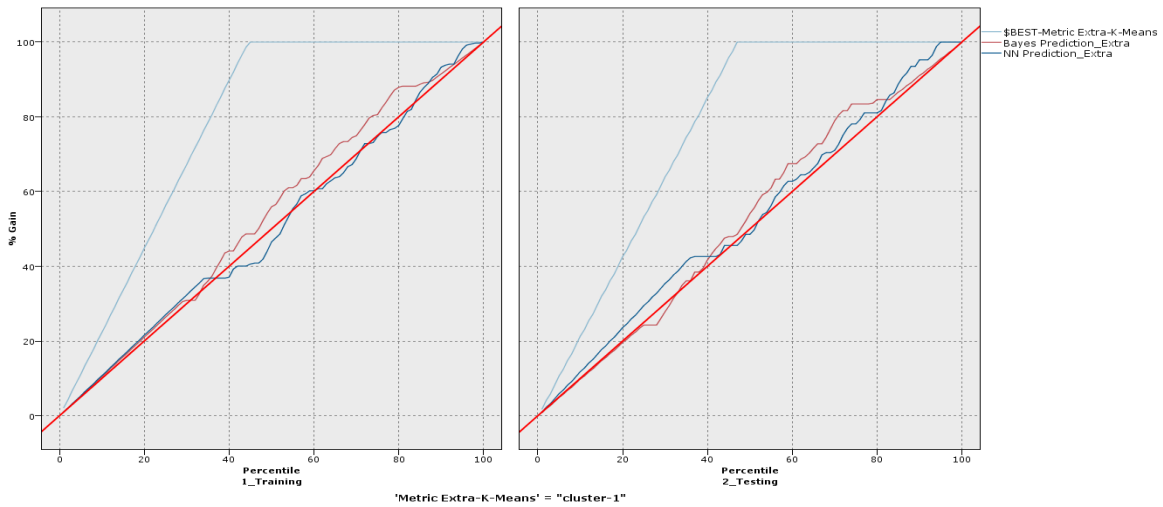


(c)

Şekil 5.28. (devam) Şehir içi yakıt tüketimini etkileyen (a) sinir ağı (b) Bayes ağı (c) C5.0 algoritması ile bulunan en önemli faktörler

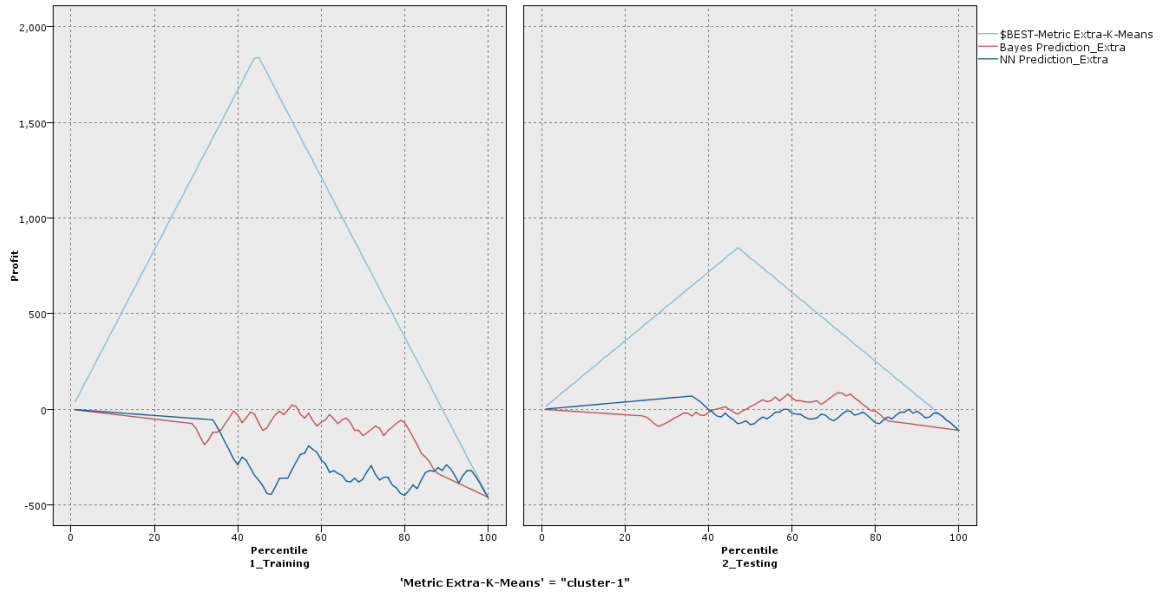
5.4.2. Şehir dışı yakıt tüketimi

Şehir dışı yakıt tüketimi verileriyle yapılan analizde ideal çözümün, verilerin %45'inden itibaren %100 kazanç ve fayda verebileceği bulunmuştur. Şekil 5.29'da görüldüğü gibi %35-88 arasında ideal çözüme en yakın kazanç ve faydayı veren teknik Bayes ağı olmuştur. Ancak, öğrenme sonrasında kazanç ve fayda açısından sinir ağı daha iyi bir tahmin vermektedir. Çözüm kalitesinin düşük olmasından dolayı C5.0 algoritması analizde görünmemektedir.



(a)

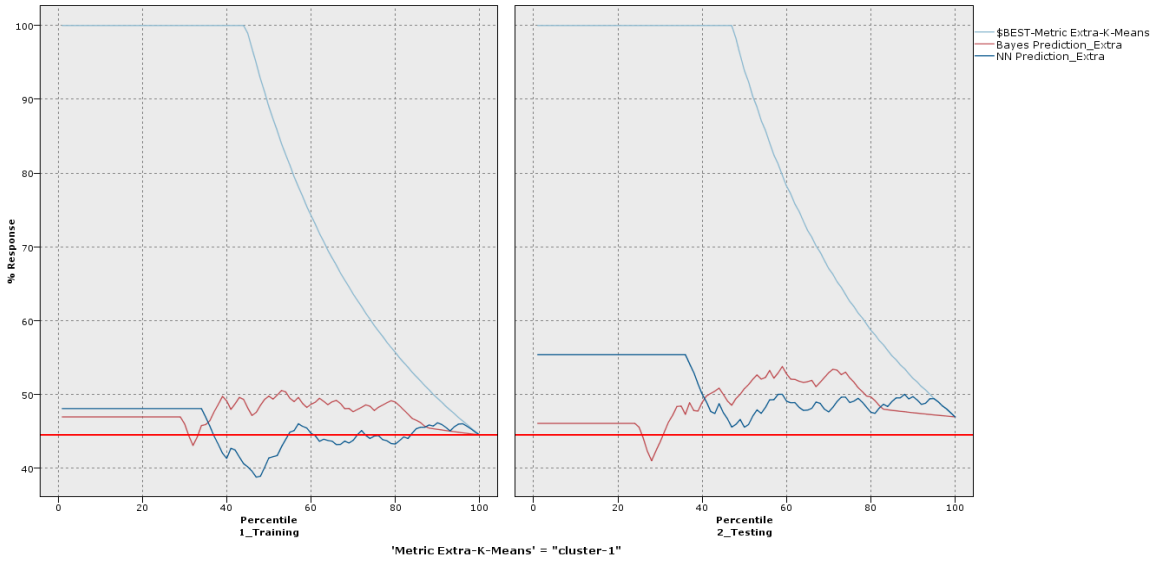
Şekil 5.29. Şehir dışı yakıt tüketimi için (a) kazanç (b) kâr grafiği



(b)

Şekil 5.29. (devam) Şehir dışı yakıt tüketimi için (a) kazanç (b) kâr grafiği

Öğrenmeyle birlikte, yükseltme, tepki ve ROI grafiklerinde ideal çözümün başarısı Şekil 5.30 ve Ek-5'teki gibi düşmüştür. Öğrenme esnasında (özellikle verinin %35-88'lik kısmı arasında) sinir ağı diğer tekniklerden daha iyi bir performans sergilemiştir. Öğrenmeden sonra da sinir ağı şehir dışı yakıt tüketiminin etkileyen en önemli faktörlerin tahmininde belirgin bir artış göstermiş ve daha iyi sonuçlar alınmıştır.



Şekil 5.30. Şehir dışı yakıt tüketimi için tepki grafiği

Şehir dışı yakıt tüketimi verileri üzerinde etkili faktörlerin belirlenmesinde sinir ağı %99, Bayes ağı %68 başarı göstermiştir (Şekil 5.31).

Comparing Bayes with Metric Extra-K-Means

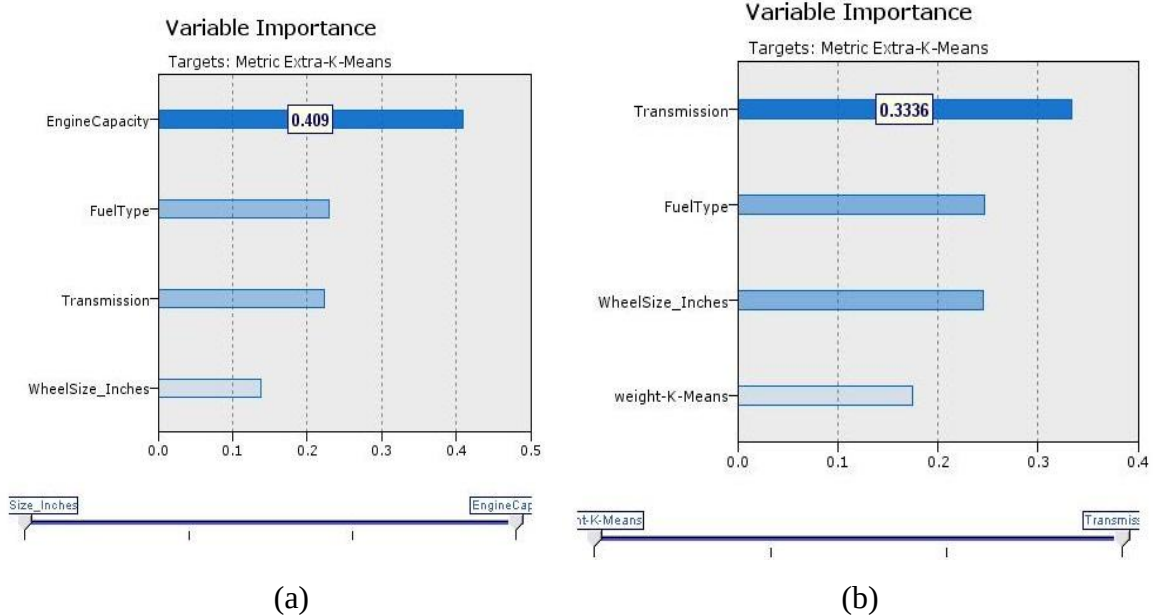
'Partition'	1_Training	2_Testing
Correct	637 76.2%	245 68.06%
Wrong	199 23.8%	115 31.94%
Total	836	360

Comparing NN with Metric Extra-K-Means

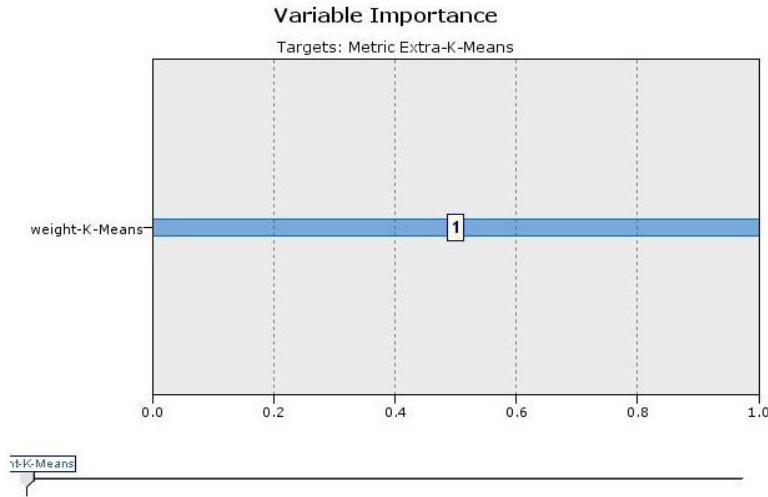
'Partition'	1_Training	2_Testing
Correct	835 99.88%	359 99.72%
Wrong	1 0.12%	1 0.28%
Total	836	360

Şekil 5.31. Şehir dışı yakıt tüketimi verilerine uygulanan tekniklerin analiz sonuçları

Sinir ağı ve Bayes ağı şehir dışı yakıt tüketimini etkileyen Şekil 5.32'deki dört parametreyi tahmin etmiştir. Sinir ağına göre vites kutusu, Bayes ağına göre ise motor kapasitesi en önemli değişkendir. Sinir ağına göre yakıt tipi ve tekerlek boyutu hemen hemen aynı önemi göstermiştir ve ardından araç ağırlığı gelmektedir. C5.0 algoritması ise sadece araç ağırlığını önemli bulmuştur. Çözümün fizibilitesi düşük olduğundan C5.0 algoritması için tahmin sonuçları dikkate alınmamıştır.



Şekil 5.32. Şehir dışı yakıt tüketimini etkileyen (a) sinir ağı (b) Bayes ağı (c) C5.0 algoritması ile bulunan en önemli faktörler

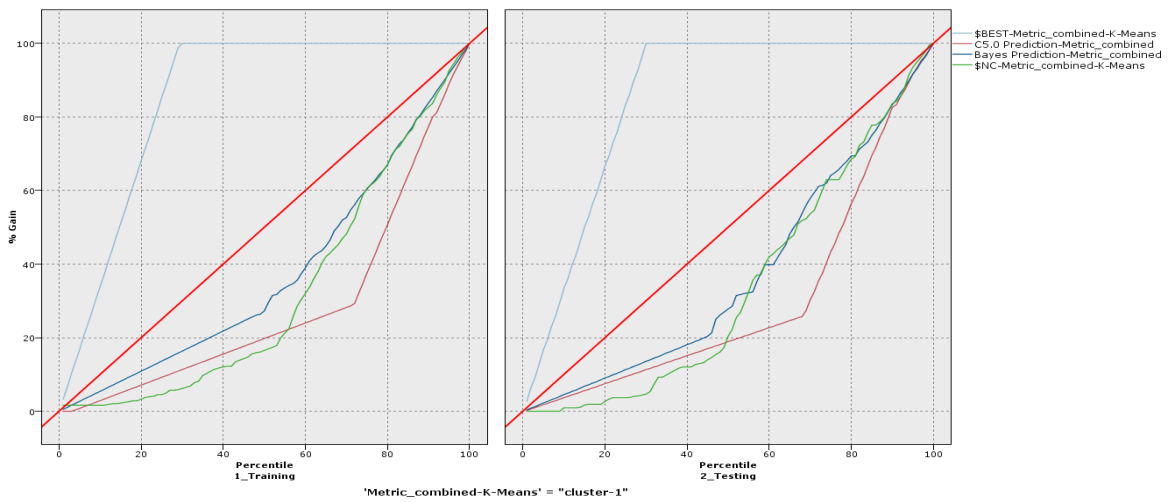


(c)

Şekil 5.32. (devam) Şehir dışı yakıt tüketimini etkileyen (a) sinir ağı (b) Bayes ağı (c) C5.0 algoritması ile bulunan en önemli faktörler

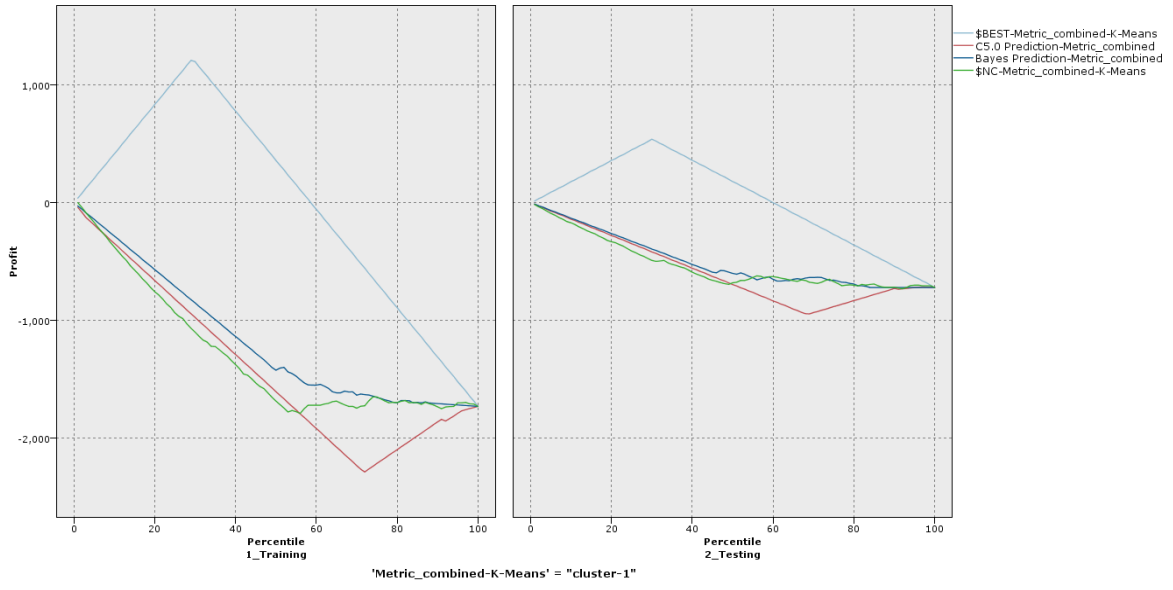
5.4.3. Kombine yakıt tüketimi

Kombine yakıt tüketimi için belirlenecek en önemli faktörler verinin %30'undan itibaren %100 kazanç ve fayda ile tahmin edilebilmektedir (Şekil 5.33). Bu ideal çözüme en yakın sonucu veren teknik Bayes ağı olmuştur. Ancak özellikle %50'den itibaren sinir ağı ile Bayes ağı hemen hemen aynı sonucu vermektedir Öğrenme, elde edilen sonuçları ideal çözüme yaklaştırsa da teknikler üzerinde büyük bir değişiklik yaratmamıştır.



(a)

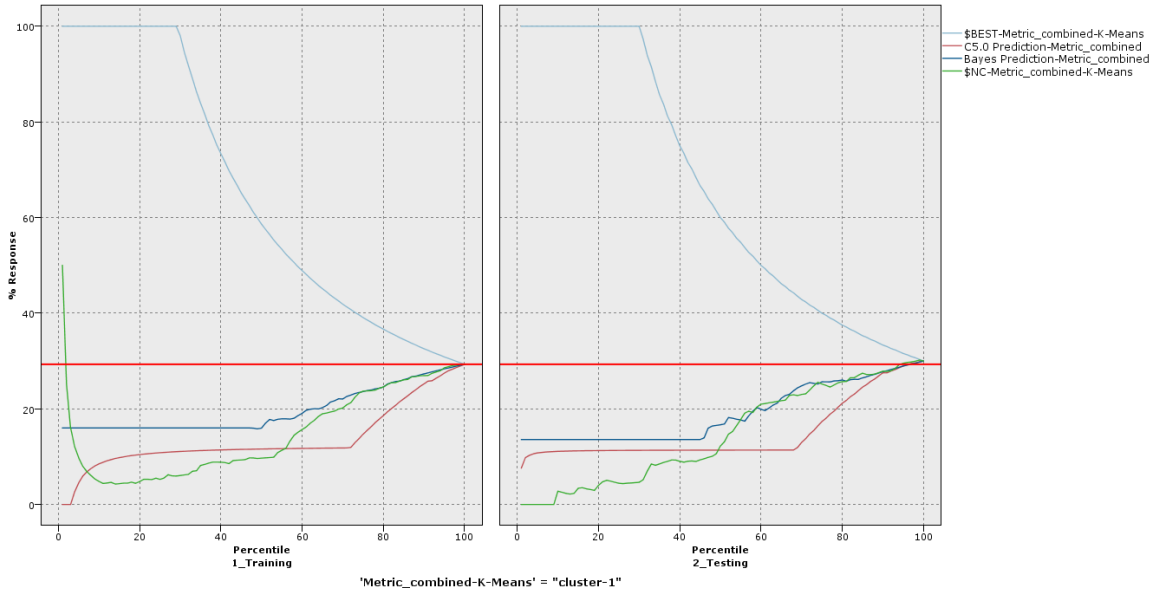
Şekil 5.33. Kombine yakıt tüketimi için (a) kazanç (b) kâr grafiği



(b)

Şekil 5.33. (devam) Kombine yakıt tüketimi için (a)kazanç (b) kâr grafiği

Kombine yakıt tüketimi verilerine uygulanan tekniklerin performansı yükseltme, tepki ve ROI grafiklerinde genellikle azalmıştır (Şekil 5.34 ve EK-6). Tahminde ideal çözüme en yakın sonuçları veren teknik Bayes ağı olmuştur ve diğer tekniklerden daha iyi bir performans göstermiştir.



Şekil 5.34. Kombine yakıt tüketimi için tepki grafiği

Önceki grafiklerde gösterildiği gibi analiz sonuçlarında da Bayes ağı ve sinir ağının gösterdiği başarı düşmüştür (Şekil 5.35). Grafiklerde en iyi çözümü Bayes ağı vermesine rağmen analiz sonuçlarına göre C5.0 algoritması daha güvenilir sonuçlar vermektedir.

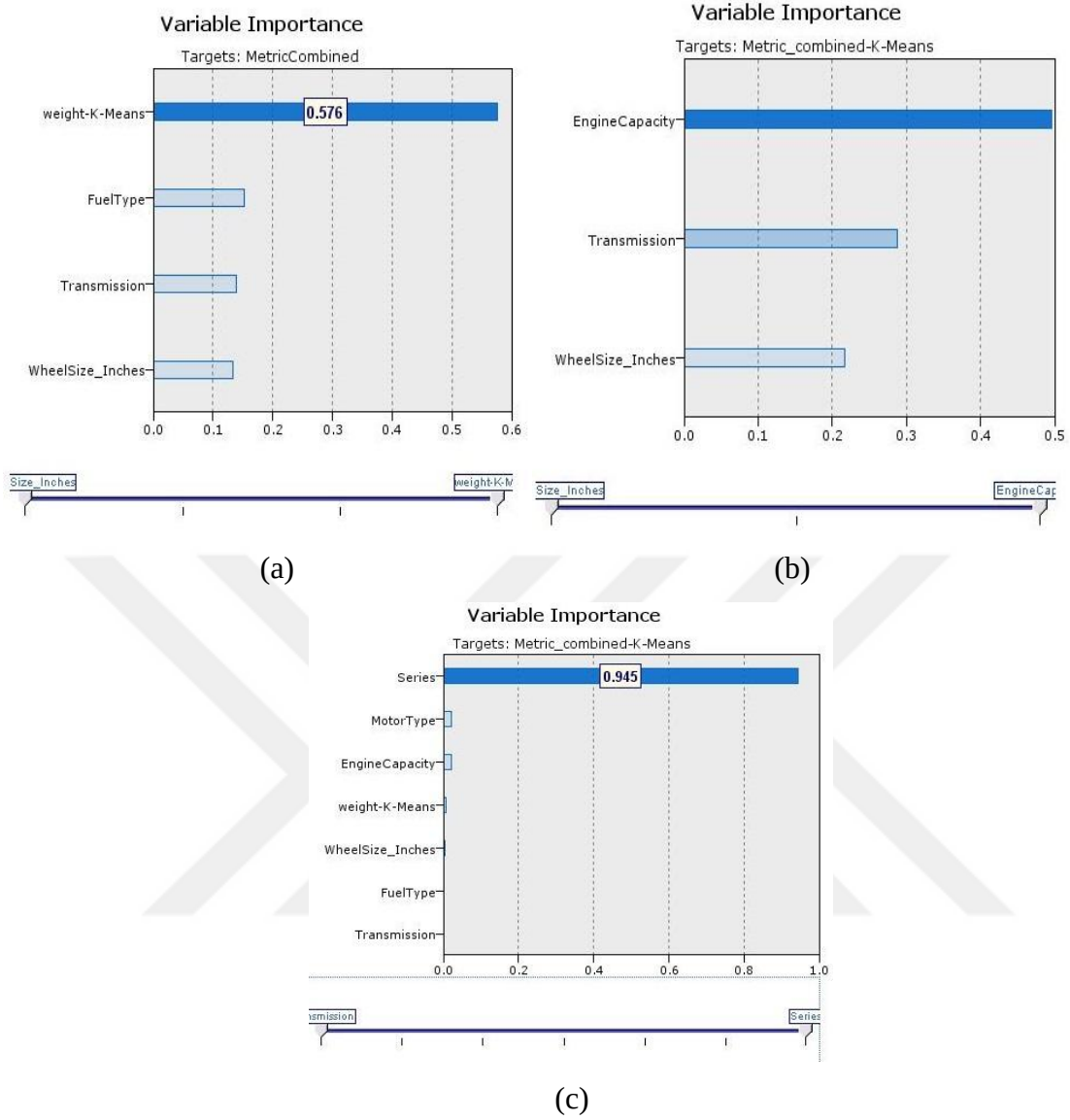
Comparing \$N\$-Metric_combined-K-Means with Metric_combined-K-Means				
'Partition'	1_Training		2_Testing	
Correct	663	79.31%	285	79.17%
Wrong	173	20.69%	75	20.83%
Total	836		360	

Comparing C5.0-Metric_combined with Metric_combined-K-Means				
'Partition'	1_Training		2_Testing	
Correct	701	83.85%	300	83.33%
Wrong	135	16.15%	60	16.67%
Total	836		360	

Comparing Bayes-Metric_combined with Metric_combined-K-Means				
'Partition'	1_Training		2_Testing	
Correct	626	74.88%	246	68.33%
Wrong	210	25.12%	114	31.67%
Total	836		360	

Şekil 5.35. Kombine yakıt tüketimi verilerine uygulanan tekniklerin analiz sonuçları

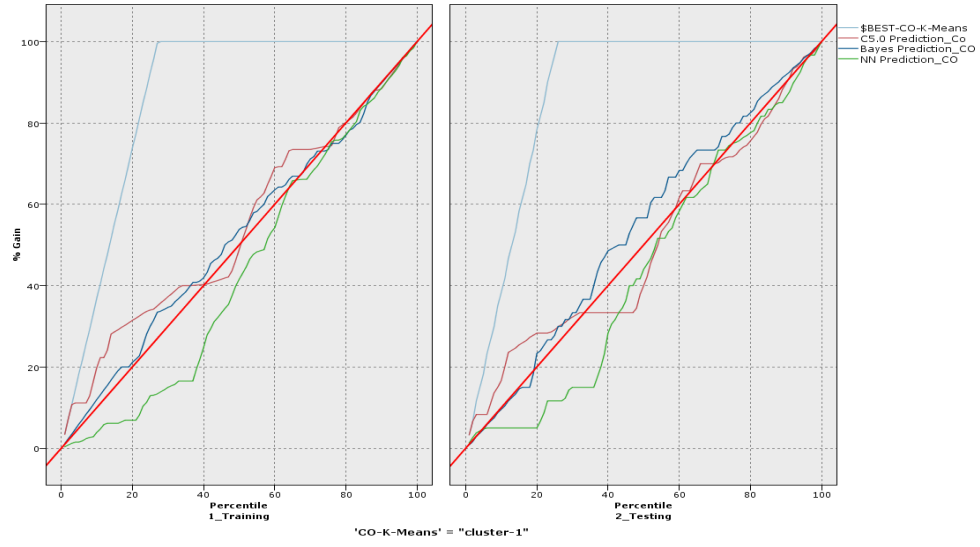
Her teknikle bulunan önemli faktörler ve sayıları birbirinden farklıdır. Sinir ağı kombine yakıt tüketimi üzerindeki en önemli faktörleri araç ağırlığı, yakıt tipi, vites kutusu ve tekerlek boyutu olarak bulmuştur (Şekil 5.36a). Bayes ağıyla bulunan faktörler ise sırasıyla motor kapasitesi, vites kutusu ve tekerlek boyutudur (Şekil 5.36b). C5.0 algoritması diğer tekniklerden biraz daha farklı sonuç göstermiştir (Şekil 5.36c). C5.0 algoritmasına göre kombine yakıt tüketimini etkileyen 7 önemli faktör vardır ve bunlardan en etkilisi otomobilin serisidir. Bunun dışında; motor tipi, motor kapasitesi, araç ağırlığı, tekerlek çapı, yakıt tipi ve vites kutusu da çok düşük oranlarda etkilidir.



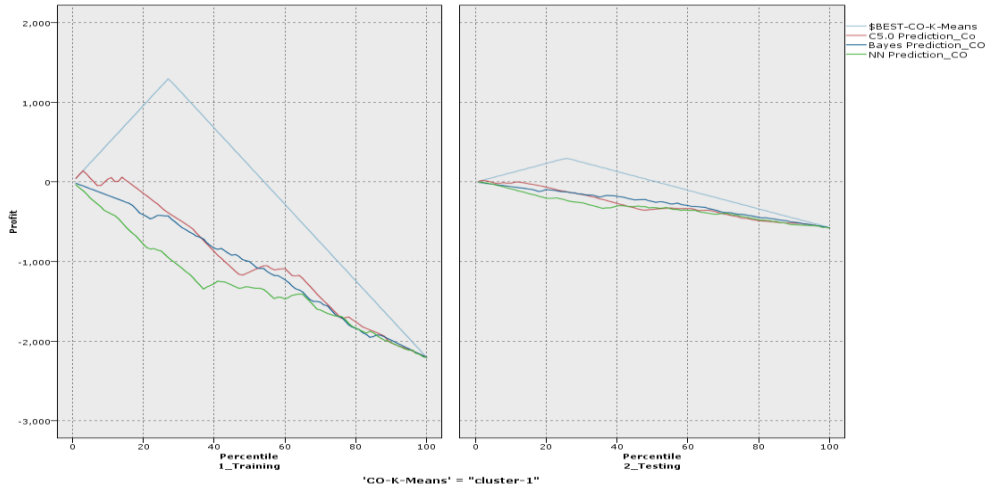
Şekil 5.36. Kombine yakıt tüketimini etkileyen (a) sinir ağı (b) Bayes ağı (c) C5.0 algoritması ile bulunan en önemli faktörler

5.4.4. Karbon monoksit (CO) salınımı

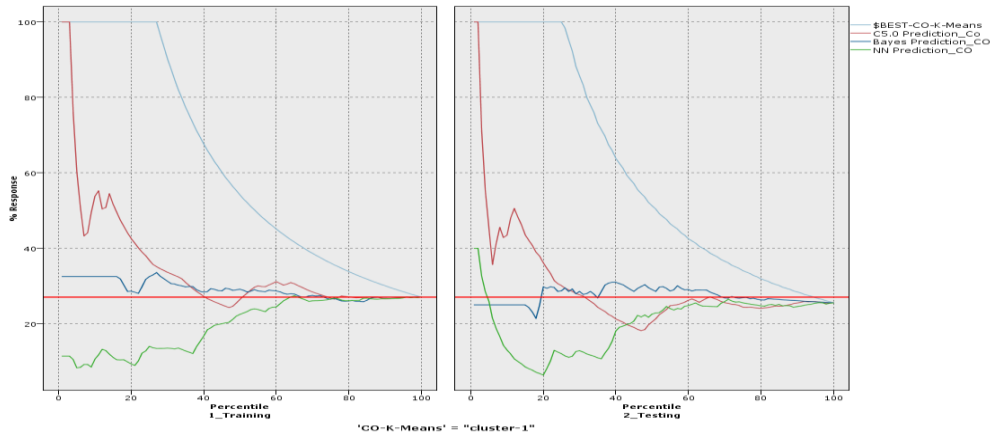
Maksimum kazanç ve fayda karbon monoksit verisinin %25'inden itibaren elde edilebilmektedir. Şekil 5.37'de, C5.0 algoritmasının sinir ağı ve Bayes ağından daha iyi tahmin yaptığı, ideal çözüme yakınlığından anlaşılmaktadır. Öğrenme sonrasında ise durum değişerek, sinir ağı karbon monoksiti etkileyen faktörler konusunda diğer tekniklerden daha iyi bir tahminde bulunmuştur.



(a)



(b)



(c)

Şekil 5.37. Karbon monoksit salınımı için (a) kazanç (b) kâr (c) yanıt grafiği

Benzer şekilde, karbon monoksit emisyonu için oluşturulan yükseltme, tepki ve ROI grafiklerinde C5.0 algoritmasının performansı öğrenmeden önce diğer tekniklerden daha üstündür (Şekil 5.37 ve EK-7). Öğrenen sinir ağları daha sonradan ideal çözüme C5.0 algoritması ve Bayes ağından daha yakın tahmin sonucu vermiştir.

Grafiklerde sinir ağının performansı yanıltıcı olmuş, analizde en güvenilir yöntem C5.0 algoritması olarak bulunmuştur (Şekil 5.38). Tekniklerin karbon monoksit salınımında en etkili faktörleri bulmadaki başarısı sırasıyla C5.0 algoritması için %72, sinir ağı için %53 ve Bayes ağı için %48 olmuştur.

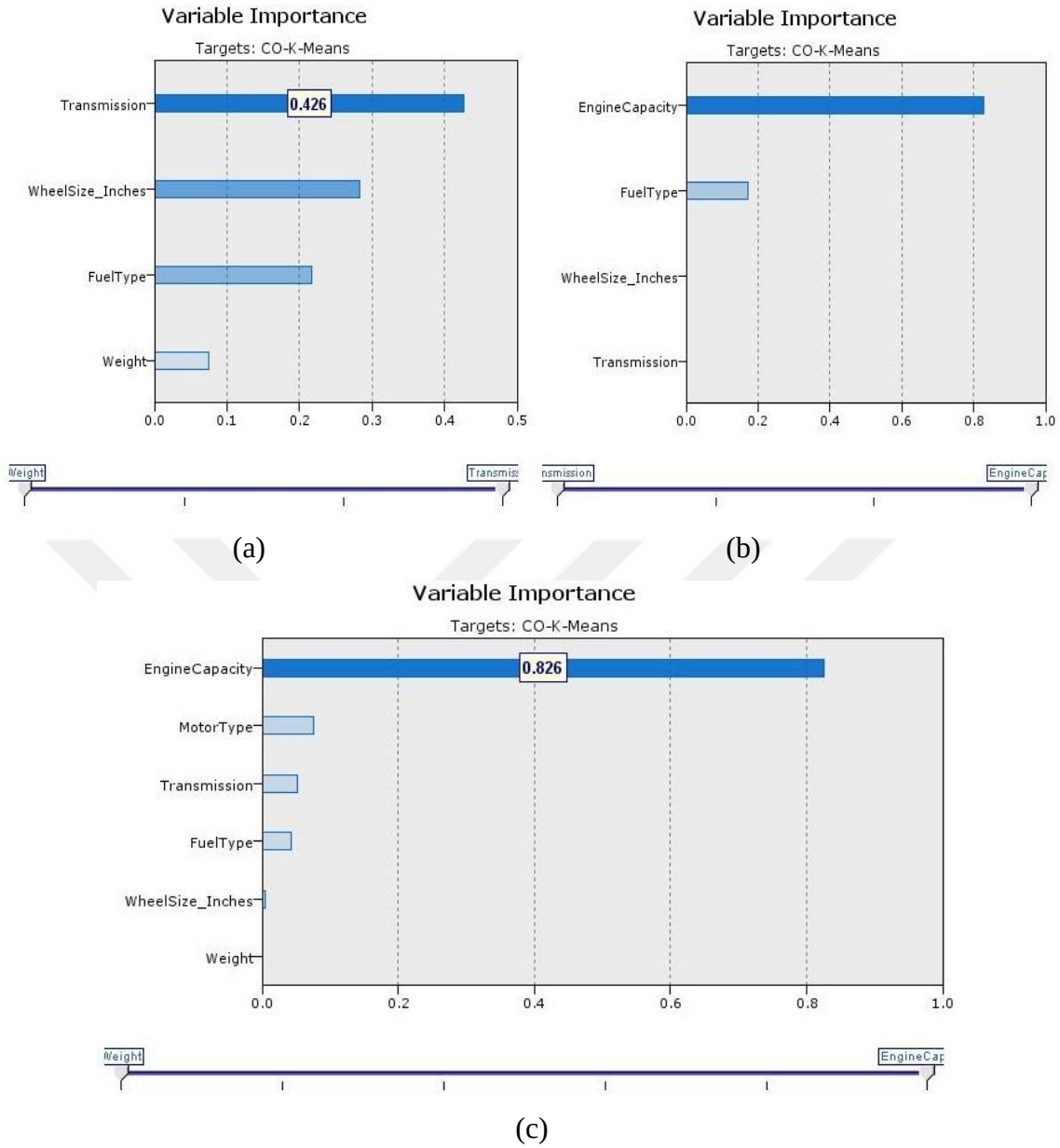
Comparing NN_CO with CO-K-Means				
'Partition'	1_Training		2_Testing	
Correct	535	55.67%	125	53.19%
Wrong	426	44.33%	110	46.81%
Total	961		235	

Comparing C5.0_CO with CO-K-Means				
'Partition'	1_Training		2_Testing	
Correct	714	74.3%	169	71.91%
Wrong	247	25.7%	66	28.09%
Total	961		235	

Comparing Bayes_CO with CO-K-Means				
'Partition'	1_Training		2_Testing	
Correct	556	57.86%	113	48.09%
Wrong	405	42.14%	122	51.91%
Total	961		235	

Şekil 5.38. Karbon monoksit (CO) verilerine uygulanan tekniklerin analiz sonuçları

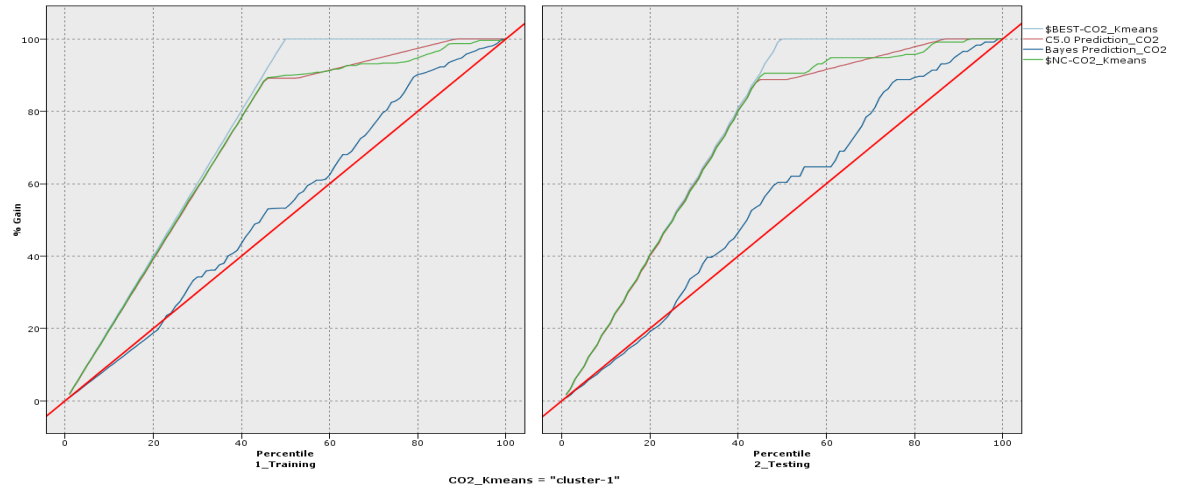
Tüm tekniklerde araçların yakıt tipi ve tekerlek boyutu önemli faktörler arasında yer almıştır (Şekil 5.39). Bayes ağı ve C5.0, karbon monoksit salınımı üzerinde motor kapasitesinin en önemli parametre olduğunu bulmuştur. Sinir ağına göre ise en önemli parametre vites kutusudur. Belirlenen en önemli faktörler sinir ağı için vites kutusu, tekerlek çapı, yakıt tipi ve araç ağırlığıdır (Şekil 5.39a). Bayes ağında motor kapasitesi, yakıt tipi, tekerlek çapı ve vites kutusu etkili bulunmuştur (Şekil 5.39b). C5.0 algoritması ise en önemli faktörleri motor kapasitesi, motor tipi, vites kutusu, yakıt tipi, tekerlek çapı ve araç ağırlığı olarak tahmin etmiştir (Şekil 5.39c).



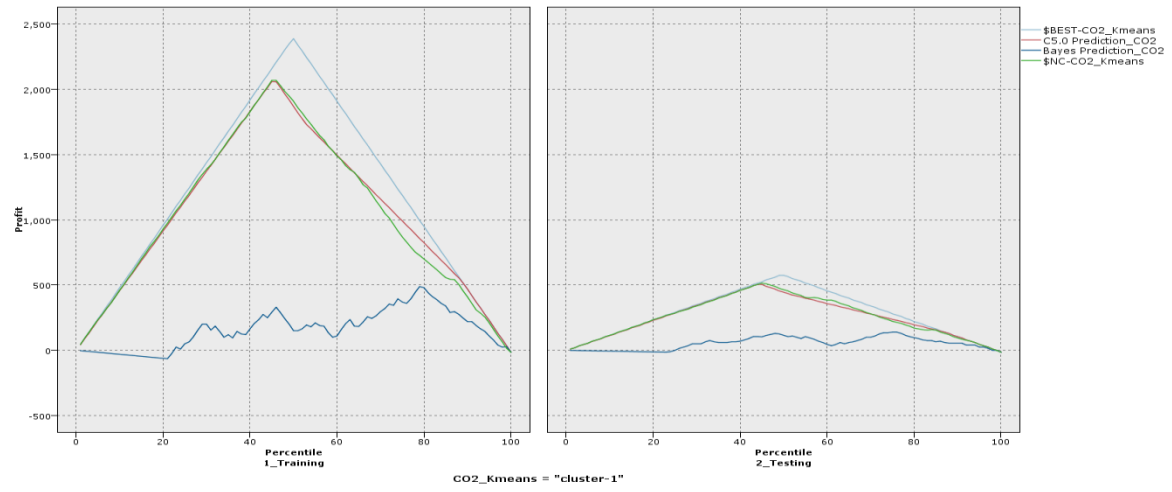
Şekil 5.39. Karbon monoksit (CO) salınımını etkileyen (a) sinir ağı (b) Bayes ağı (c) C5.0 algoritması ile bulunan en önemli faktörler

5.4.5. Karbondioksit (CO₂) salınımı

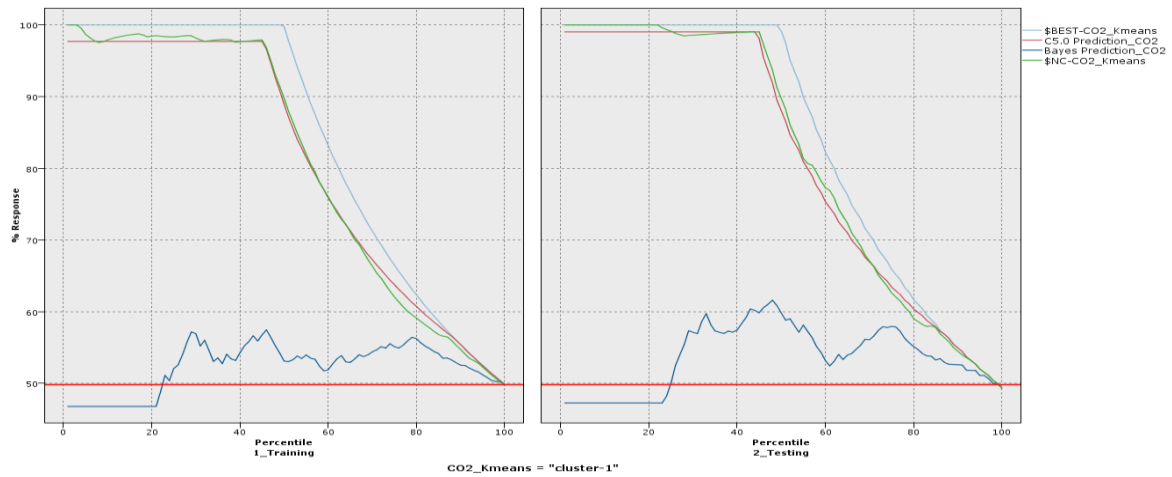
Karbondioksit verileri üzerinde yapılan analizde ideal çözüm verinin %50'sinden itibaren %100 kazanç ve fayda sağlamaktadır. Şekil 5.40'ta görüldüğü gibi, %2-45 ve %85-100 aralığında sinir ağı ile C5.0 algoritması hemen hemen aynı tahminleri yapmıştır. Öğrenmenin ardından Bayes ağının kazanç ve fayda değerleri C5.0 algoritmasını geride bırakmıştır.



(a)



(b)



(c)

Şekil 5.40. Karbondioksit (CO_2) salınımı için (a)kazanç (b) kâr (c) yanıt grafiği

Aynı şekilde tepki, yükseltme ve ROI grafiklerinde de Bayes ağı ile C5.0 algoritması benzer tahmin göstermiştir (Şekil 5.40 ve EK-8). Öğrenme sonrasında sinir ağının performansı da artmıştır. Yine de karbon dioksit salınımı üzerinde etkili faktörleri belirlemede en iyi yöntem Bayes ağı olmuştur.

Analiz sonuçlarında da Bayes ağı %91.5 başarı göstererek karbon dioksit üzerinde etkili parametreleri belirlemede en iyi performansı göstermiştir (Şekil 5.41).

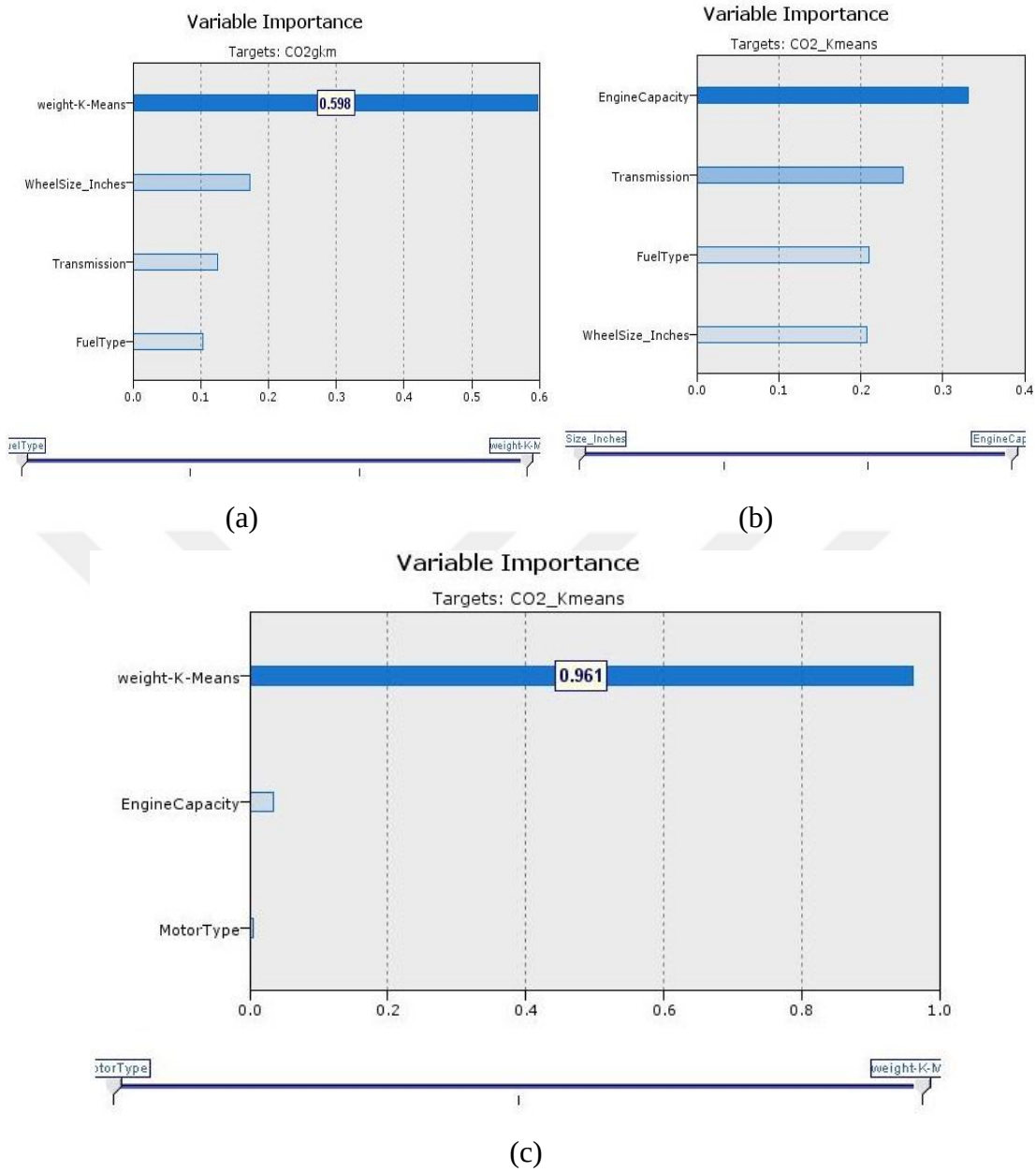
Comparing \$N_CO2_Kmeans with CO2_Kmeans					
'Partition'	1_Training		2_Testing		
Correct	870	90.53%	212	90.21%	
Wrong	91	9.47%	23	9.79%	
Total	961		235		

Comparing C5.0_CO2 with CO2_Kmeans					
'Partition'	1_Training		2_Testing		
Correct	873	90.84%	212	90.21%	
Wrong	88	9.16%	23	9.79%	
Total	961		235		

Comparing Bayes_CO2 with CO2_Kmeans					
'Partition'	1_Training		2_Testing		
Correct	881	91.68%	215	91.49%	
Wrong	80	8.32%	20	8.51%	
Total	961		235		

Şekil 5.41. Karbon dioksit (CO) salınımı verilerine uygulanan tekniklerin analiz sonuçları

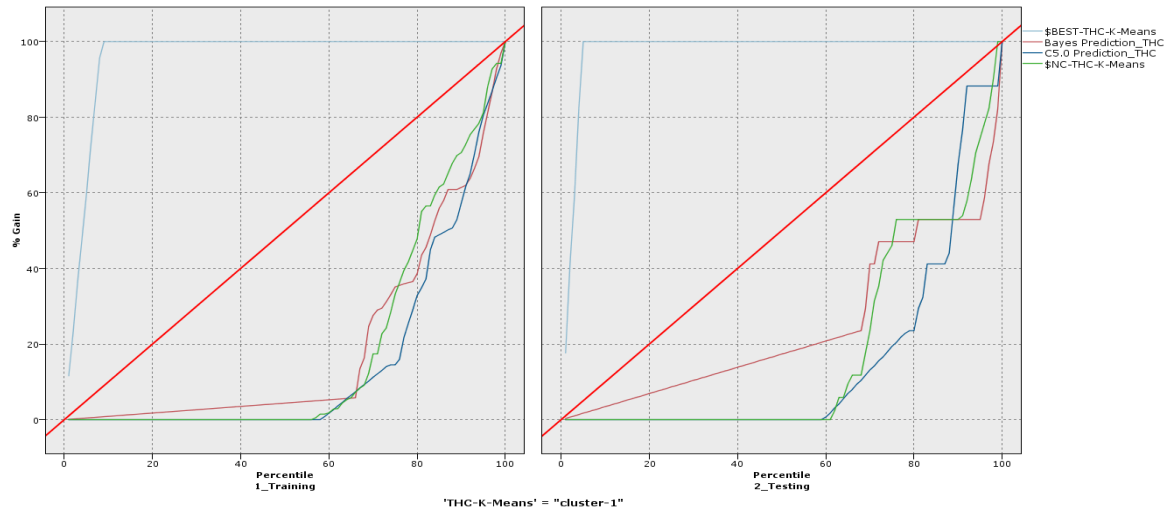
Hem sinir ağı hem de C5.0 algoritması, araç ağırlığını karbon dioksit salınımını etkileyen en önemli faktör olarak bulmuştur (Şekil 5.42). Sinir ağına göre sırasıyla tekerlek boyutu, vites kutusu ve yakıt tipi en önemli diğer faktörlerdir (Şekil 5.42a). Bayes ağı bu sırayı motor kapasitesi, vites kutusu, yakıt tipi ve tekerlek boyutu olarak bulmuştur (Şekil 5.42b). C5.0 algoritmasına göre en önemli faktörler ağırlık, motor kapasitesi ve motor tipidir (Şekil 5.42c).



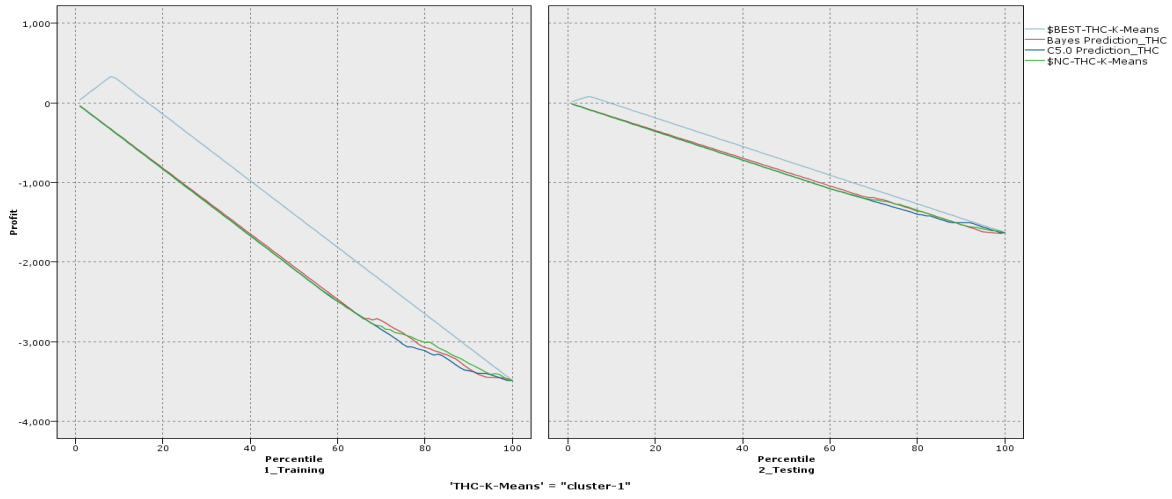
Şekil 5.42. Karbon dioksit (CO₂) salınımını etkileyen (a) sinir ağı (b) Bayes ağı (c) C5.0 algoritması ile bulunan en önemli faktörler

5.4.6. Hidrokarbon (THC) emisyonu

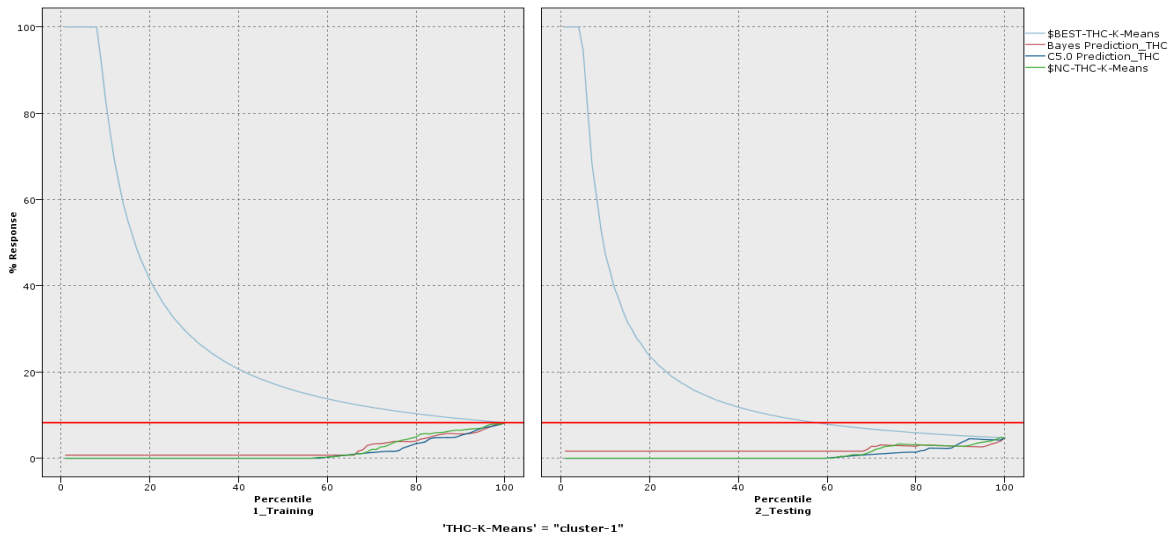
Analizlerde ideal olarak verinin %10'undan itibaren %100 kazanç elde etmek mümkündür (Şekil 5.43). Ancak, benzer fayda göstermelerine rağmen tekniklerin kazançları farklılık göstermektedir. Genel olarak Bayes ağı kazanç açısından en iyi performansı göstermektedir. Fayda grafiklerinde üç teknik de hemen hemen aynı sonuçları vermiştir.



(a)



(b)



(c)

Şekil 5.43. Hidrokarbon (THC) salınımı için (a) kazanç (b) kâr (c) yanıt grafiği

Şekil 5.43 ve EK-9’da görüldüğü gibi, Bayes ağı tepki, yükseltme ve ROI grafiklerinde de diğer tekniklerden daha iyi bir tahminde bulunmuştur. Öğrenme bu sonucu değiştirmemiştir.

Ancak analiz sonuçları sinir ağının yüksek performans göstermesine rağmen C5.0 algoritmasının diğer tekniklerden daha iyi tahminde bulunduğunu göstermiştir (Şekil 5.44).

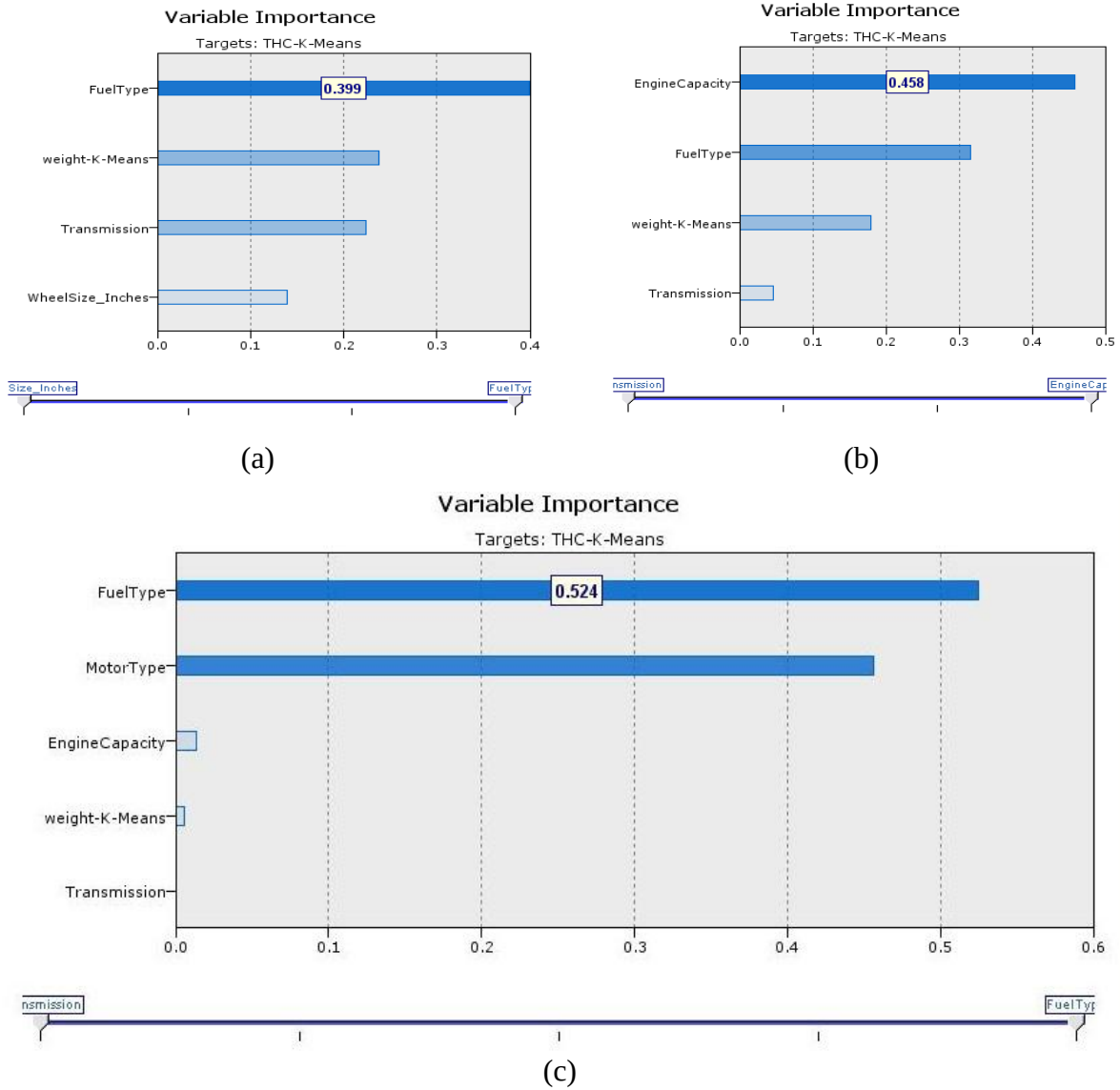
Comparing \$N\$-THC-K-Means with THC-K-Means				
'Partition'	1_Training		2_Testing	
Correct	685	81.94%	298	82.78%
Wrong	151	18.06%	62	17.22%
Total	836		360	

Comparing C5.0_THC with THC-K-Means				
'Partition'	1_Training		2_Testing	
Correct	733	87.68%	316	87.78%
Wrong	103	12.32%	44	12.22%
Total	836		360	

Comparing Bayes_THC with THC-K-Means				
'Partition'	1_Training		2_Testing	
Correct	683	81.7%	302	83.89%
Wrong	153	18.3%	58	16.11%
Total	836		360	

Şekil 5.44. Hidrokarbon (THC) emisyon verilerine uygulanan tekniklerin analiz sonuçları

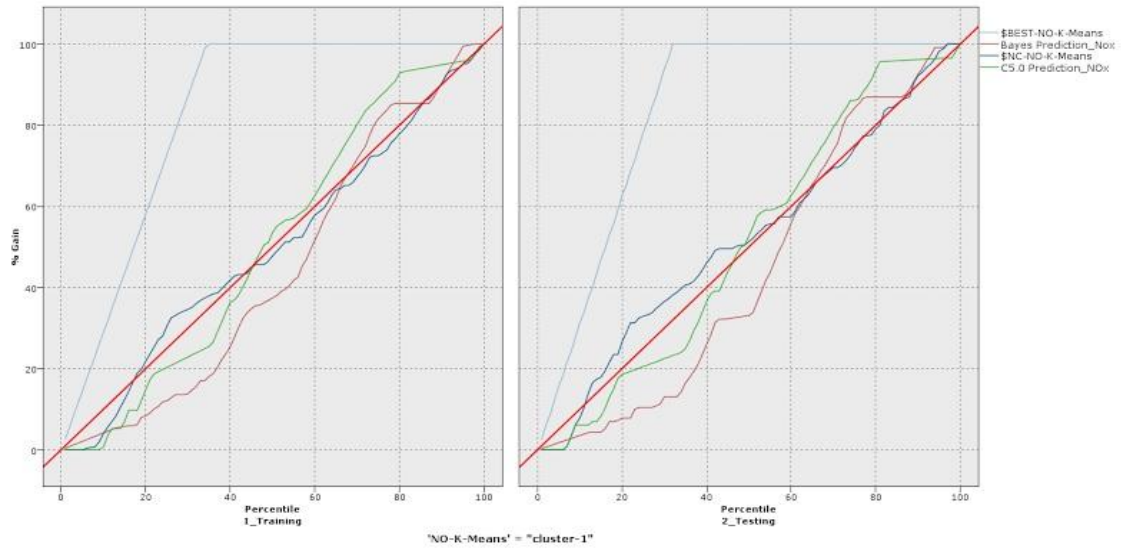
Sinir ağları ve C5.0 algoritması, hidrokarbon emisyonu üzerinde en etkili faktörü yakıt tipi, Bayes ağı ise motor kapasitesi olarak bulmuştur (Şekil 5.45). Sinir ağıyla bulunan diğer en önemli faktörler sırasıyla vites kutusu, ağırlık ve tekerlek boyutu olmuştur. C5.0 algoritması ise bundan farklı olarak motor tipini de önemli bulmuştur.



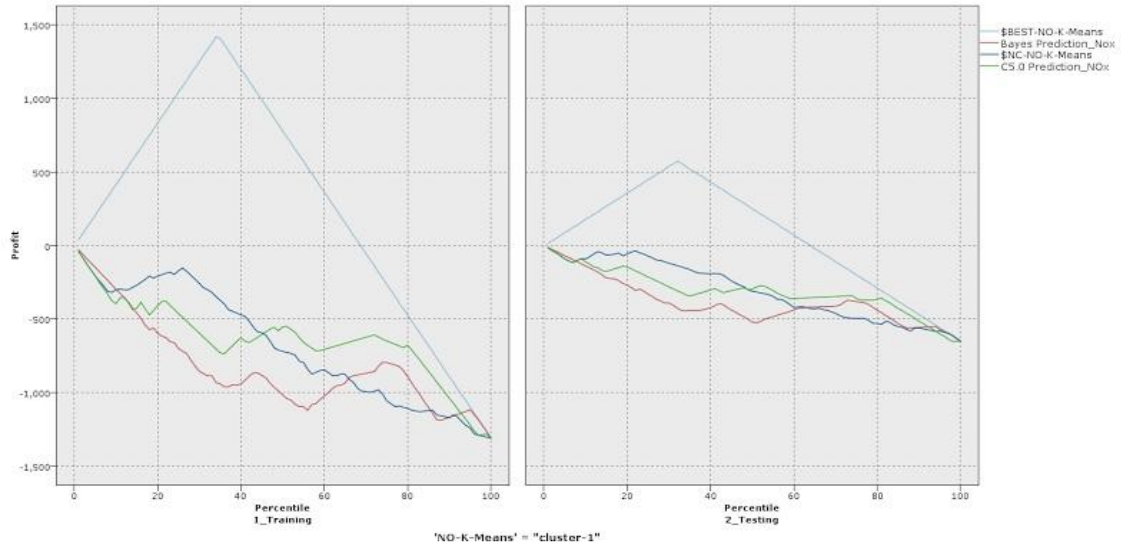
Şekil 5.45. Hidrokarbon (THC) emisyonunu etkileyen (a) sinir ağı (b) Bayes ağı (c) C5.0 algoritması ile bulunan en önemli faktörler

5.4.7. Nitrojen oksit (NO_x) emisyonu

Nitrojen oksitler üzerinde etkili faktörlerin belirlenmesinde, teknikler verinin %35'inden itibaren kazanç ve fayda açısından %100 performans gösterebilmektedir (Şekil 5.46). C5.0 algoritması ve sinir ağı Bayes ağından daha iyi performans göstermektedir. Öğrenme sonrasında da sinir ağının kazanç değeri yükselmiş ve tüm tekniklerin fayda değerleri birbirine yaklaşmıştır.



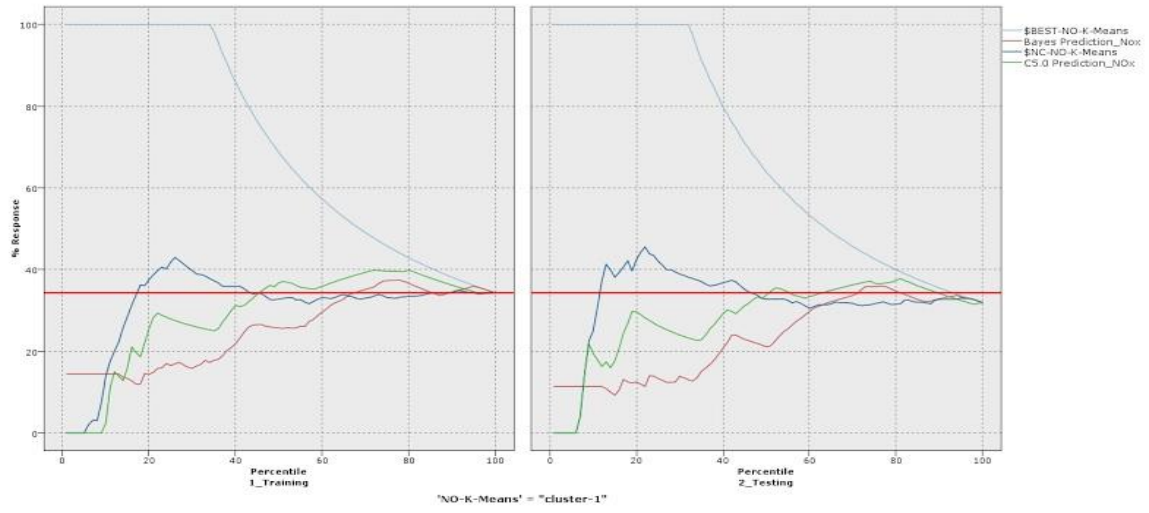
(a)



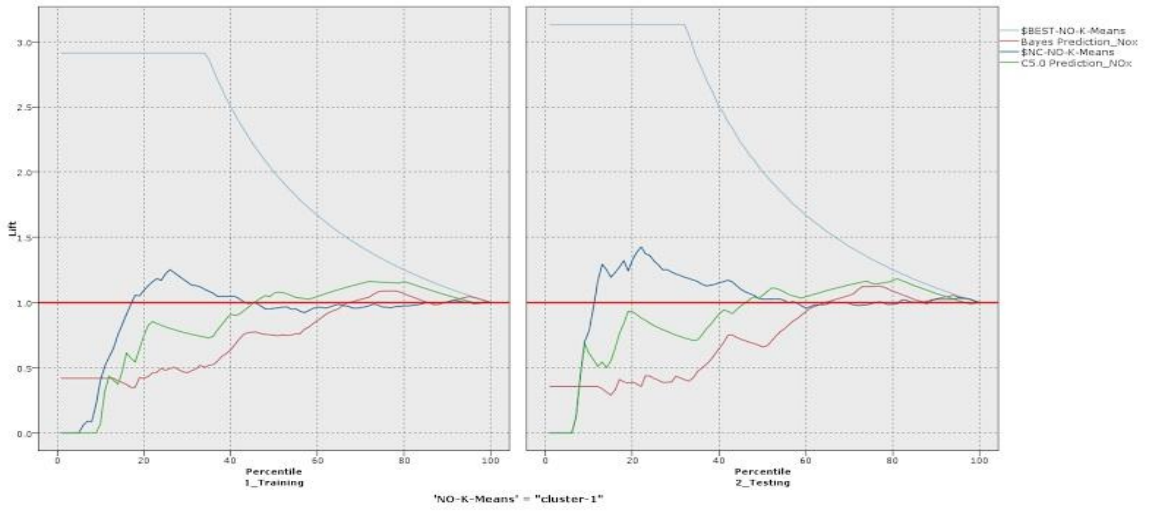
(b)

Şekil 5.46. Nitrojen oksit (NO_x) salınımı için (a) kazanç (b) kâr grafiği

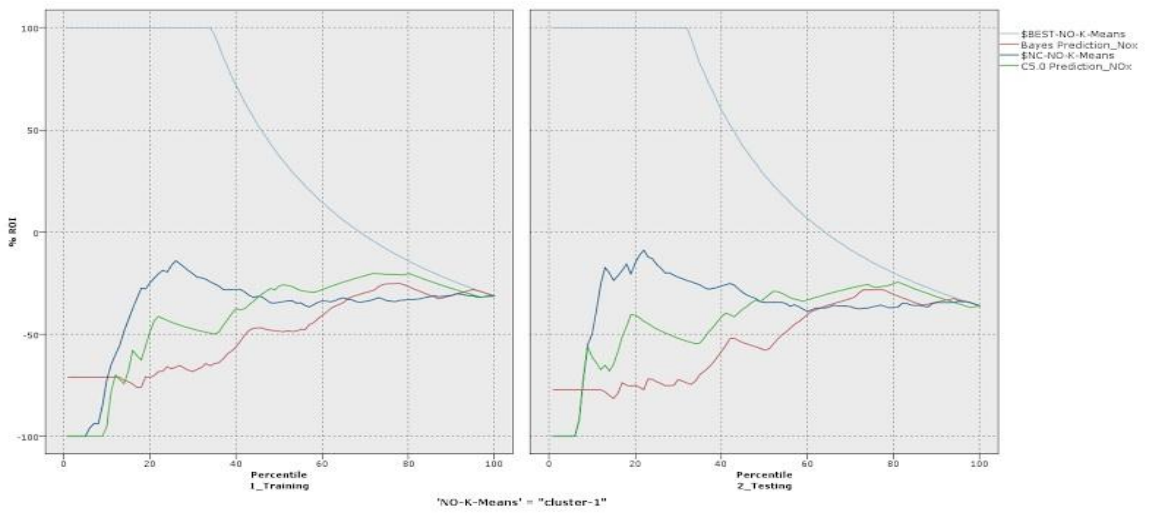
Nitrojen oksit verilerinin %0-8 ve %96-97 aralıklarında Bayes ağı en iyi teknik olmuştur (Şekil 5.47). Öğrenmeden sonra, sinir ağı verinin %8-48 ve C5.0 algoritması %48-96 aralıklarında ideal çözüme benzer sonuçlar vermektedir. %97-100 aralığında ise hem Bayes hem sinir ağı tepki, yükseltme ve ROI açısından nispeten daha iyi sonuçlar vermiştir.



(a)



(b)



(c)

Şekil 5.47. Nitrojen oksit (NO_x) emisyonu için (a) tepki, (b) yükseltme ve (c) ROI grafikleri

Nitrojen oksit verilerine uygulanan tekniklerin en önemli faktörleri tahmin etme oranları analizde sırasıyla C5.0 algoritması için %73, Bayes ağı için %66 ve sinir ağı için %53 bulunmuştur (Şekil 5.48). Çok yüksek bir performans gözlenmese de NO_x için en iyi tahmini C5.0 algoritması yapmaktadır.

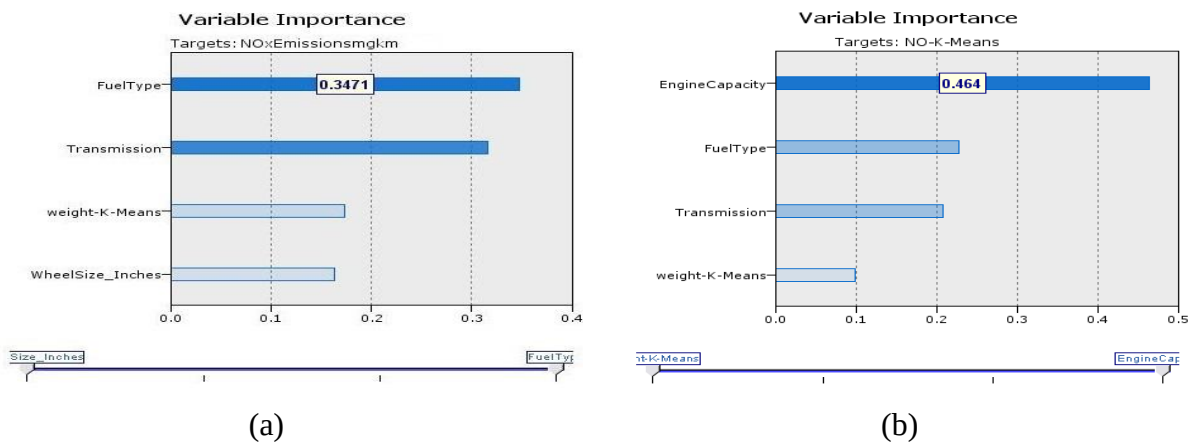
Comparing \$N\$-NO-K-Means with NO-K-Means					
'Partition'	1_Training		2_Testing		
Correct	463	55.38%	192	53.33%	
Wrong	373	44.62%	168	46.67%	
Total	836		360		

Comparing C5.0_NOx with NO-K-Means					
'Partition'	1_Training		2_Testing		
Correct	622	74.4%	263	73.06%	
Wrong	214	25.6%	97	26.94%	
Total	836		360		

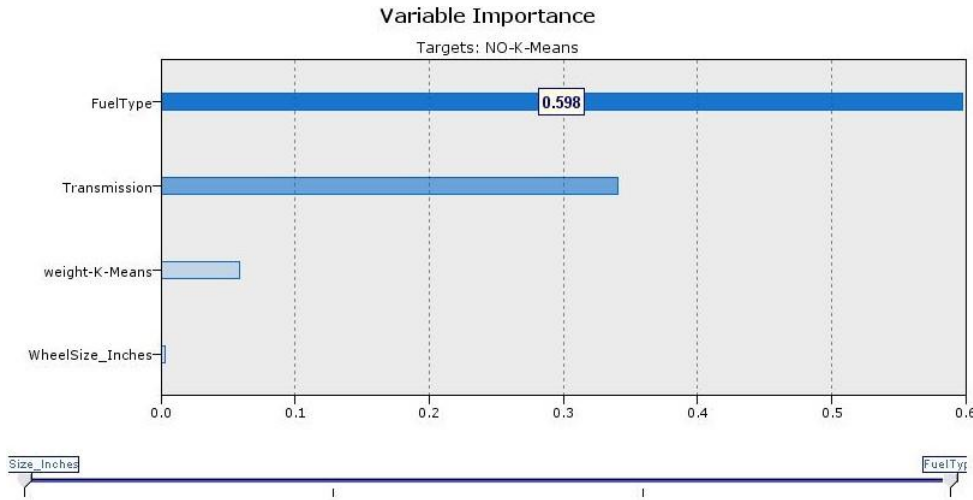
Comparing Bayes_NOx with NO-K-Means					
'Partition'	1_Training		2_Testing		
Correct	549	65.67%	237	65.83%	
Wrong	287	34.33%	123	34.17%	
Total	836		360		

Şekil 5.48. Nitrojen oksit (NO_x) emisyon verilerine uygulanan tekniklerin analiz sonuçları

Sinir ağı, nitrojen oksit emisyonu üzerinde en etkili faktörleri vites kutusu, yakıt tipi, ağırlık ve tekerlek boyutu olarak bulmuştur (Şekil 5.49a). Bayes ağına göre en önemli faktörler araçların motor kapasitesi, yakıt tipi ve vites kutusudur (Şekil 5.49b). Aynı şekilde yakıt tipi, vites kutusu ve tekerlek boyutu C5.0 algoritması ile de önemli bulunmuştur (Şekil 5.49c).



Şekil 5.49. Nitrojen Oksit (NO_x) emisyonunu etkileyen (a) sinir ağı (b) Bayes ağı (c) C5.0 algoritması ile bulunan en önemli faktörler

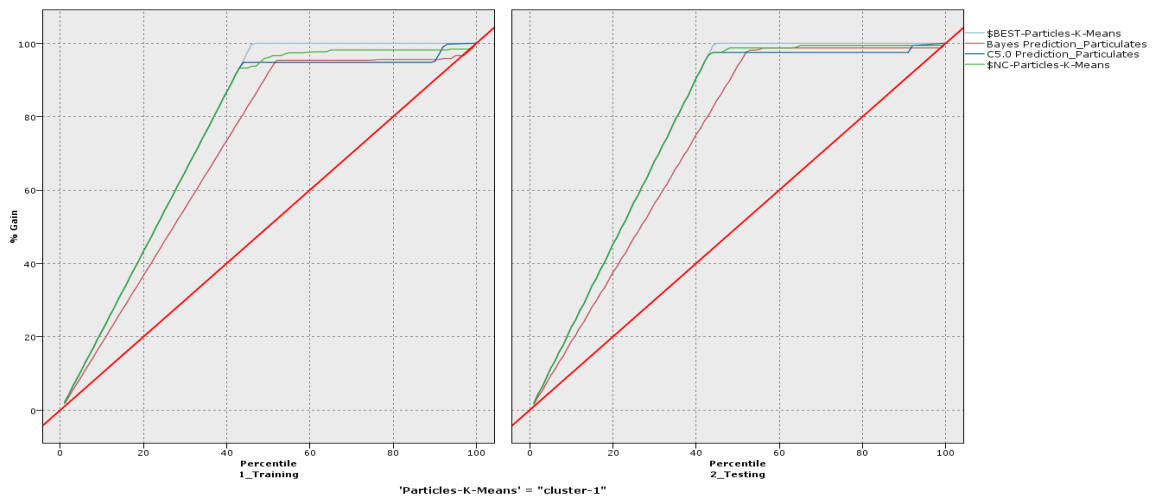


(c)

Şekil 5.49. (devam) Nitrojen Oksit (NO_x) emisyonunu etkileyen (a) sinir ağı (b) Bayes ağı (c) C5.0 algoritması ile bulunan en önemli faktörler

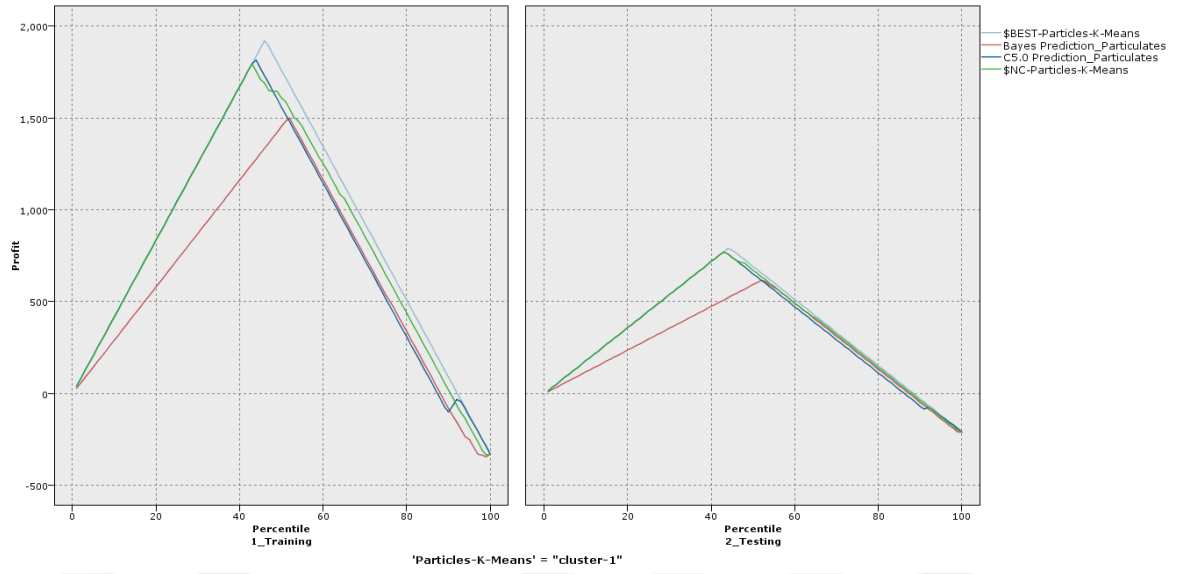
5.4.8. Parçacık emisyonu

Parçacık emisyonu üzerinde etkili olan faktörlerin belirlenmesinde verinin %45'inden itibaren %100 kazanç ve fayda elde edilebilmektedir. Öncesinde, Şekil 5.50'de görüldüğü gibi, C5.0 algoritması ve sinir ağı kazanç ve fayda açısından ideal çözüme en yakın tahminleri yapmaktadır. Öğrenmeden sonra bütün tekniklerin kazanç ve fayda değerleri yükselmiştir.

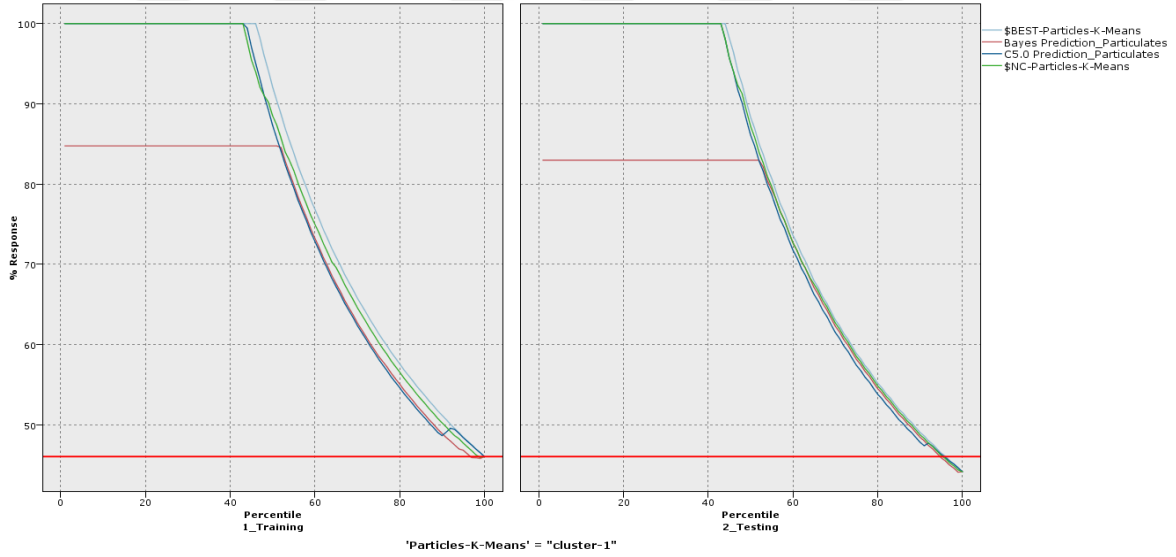


(a)

Şekil 5.50. Parçacık emisyonu için (a) kazanç (b) kâr (c) yanıt grafiği



(b)



(c)

Şekil 5.50. (devam) Parçacık emisyonu için (a) kazanç (b) kâr (c) yanıt grafiği

Tepki, yükseltme ve ROI grafiklerinde tüm teknikler ideal çözüme yakın bir tahminde bulunmuştur (Şekil 5.50 ve EK-10). Öğrenme ile bütün tekniklerin performansı artmış ve sinir ağı ile C5.0 algoritması Bayes ağından daha iyi bir tahminde bulunmuştur.

Analiz sonuçlarına göre parçacık emisyonunu etkileyen en önemli faktörlerin tahmininde C5.0 algoritması diğerlerinden daha iyi bir performans göstermiştir (Şekil 5.51). Tekniklerin performans değerleri sırasıyla C5.0 algoritması için %96, Bayes ağı için %95 ve sinir ağı için %90 olmuştur.

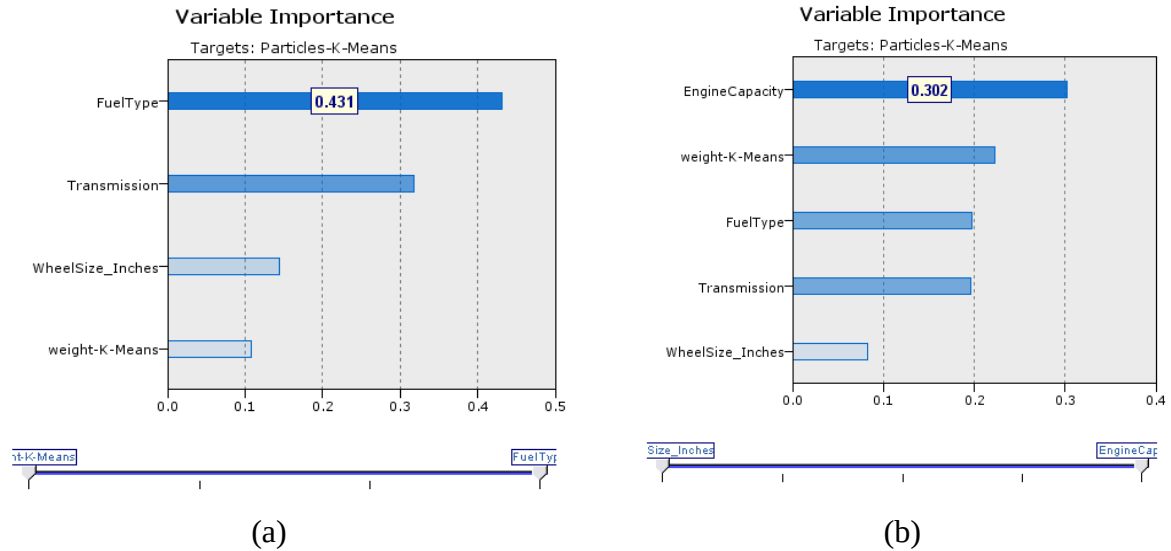
Comparing \$N\$-Particles-K-Means with Particles-K-Means				
'Partition'	1_Training		2_Testing	
Correct	747	89.35%	324	90%
Wrong	89	10.65%	36	10%
Total	836		360	

Comparing C5.0_Particates with Particles-K-Means				
'Partition'	1_Training		2_Testing	
Correct	788	94.26%	346	96.11%
Wrong	48	5.74%	14	3.89%
Total	836		360	

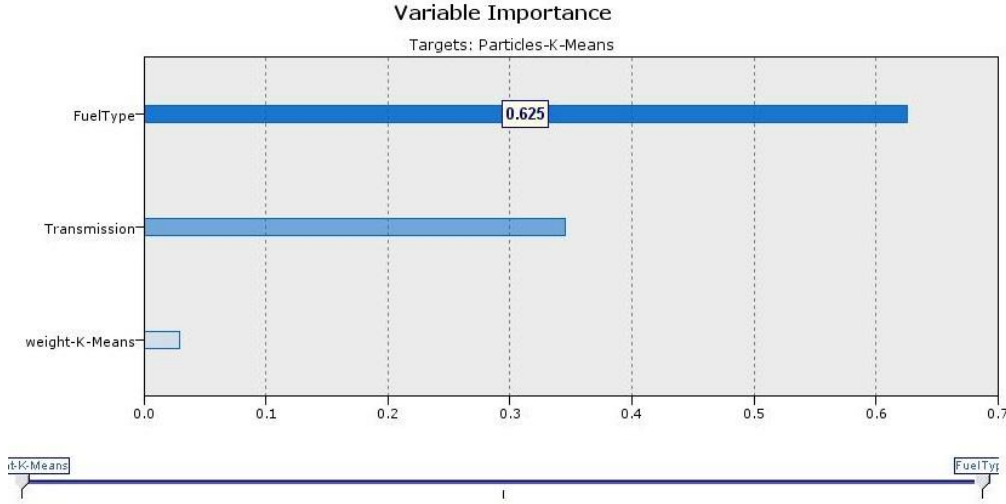
Comparing Bayes_Particates with Particles-K-Means				
'Partition'	1_Training		2_Testing	
Correct	779	93.18%	342	95%
Wrong	57	6.82%	18	5%
Total	836		360	

Şekil 5.51. Parçacık emisyonu verilerine uygulanan tekniklerin analiz sonuçları

Parçacık emisyonunu en çok etkileyen faktör sinir ağına göre yakıt tipidir (Şekil 5.52a). Diğer önemli faktörler sırasıyla vites kutusu, tekerlek boyutu ve araç ağırlığı olmuştur. Bayes ağıyla ise en önemli faktörler sırasıyla motor kapasitesi, vites kutusu, yakıt tipi ve tekerlek çapıdır (Şekil 5.52b). C5.0 algoritması da sinir ağına benzer bir sonuç vererek en önemli faktörleri sırasıyla yakıt tipi, vites kutusu ve ağırlık olarak bulmuştur (Şekil 5.52c).



Şekil 5.52. Parçacık emisyonunu etkileyen (a) sinir ağı (b) Bayes ağı (c) C5.0 algoritması ile bulunan en önemli faktörler

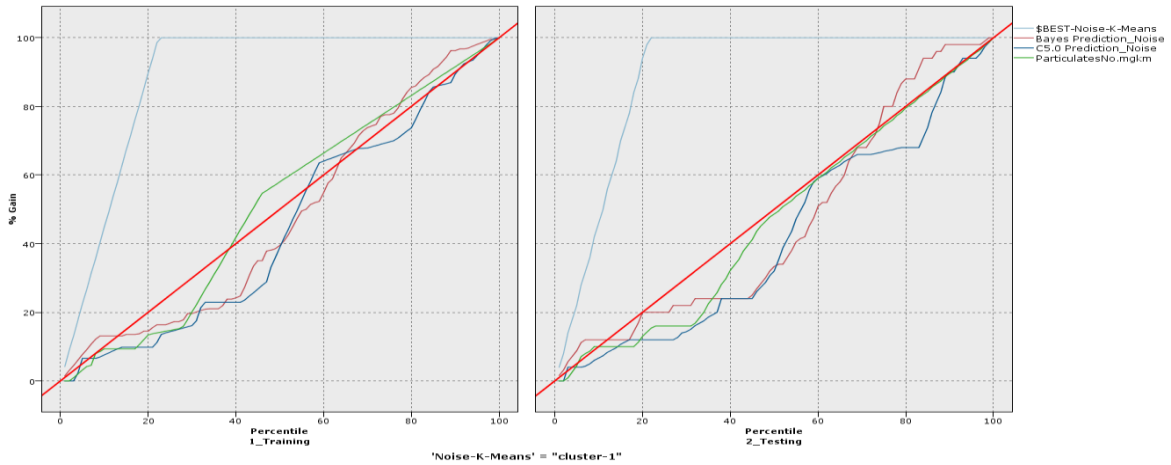


(c)

Şekil 5.52. (devam) Parçacık emisyonunu etkileyen (a) sinir ağı (b) Bayes ağı (c) C5.0 algoritması ile bulunan en önemli faktörler

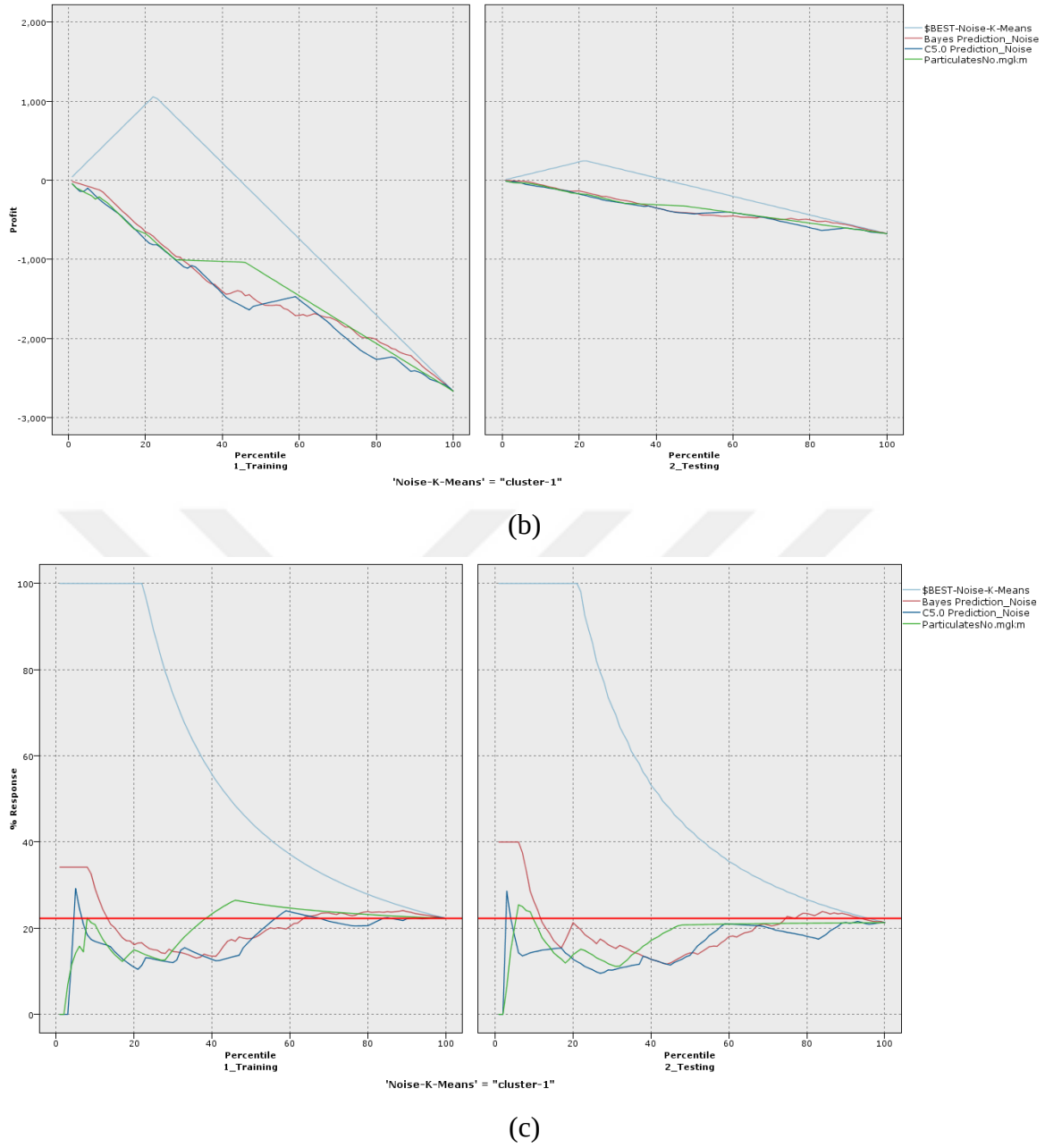
5.4.9. Gürültü seviyesi

Yaklaşık olarak gürültü seviyesi verilerinin %22'sinden itibaren ideal çözüm %100 kazanç sağlayabilmektedir (Şekil 5.53a). %0-28 ve %78-100 aralıklarında Bayes ağı ideal çözüme daha yakın sonuçlar vererek diğer tekniklerden daha iyi bir tahmin yapmıştır. %28-78 arasında sinir ağı, kazanç ve fayda açısından diğer tekniklerden daha iyi bir performans göstermektedir. Şekil 5.53b, öğrenme sonrasında tüm tekniklerin performansının yükselerek homojen sonuçlar elde edildiğini göstermektedir.



(a)

Şekil 5.53. Gürültü seviyesi için (a) kazanç (b) kâr (c) yanıt grafiği



Şekil 5.53. (devam) Gürültü seviyesi için (a) kazanç (b) kâr (c) yanıt grafiği

Gürültü seviyesini etkileyen faktörlerin tahmininde tüm teknikler tepki, yükseltme ve ROI grafiklerinde benzer sonuçlar göstermiştir (Şekil 5.53 ve EK-11).

Analiz sonuçlarında gürültü seviyesini etkileyen en önemli faktörler sırasıyla C5.0 algoritmasıyla 65,5, sinir ağıyla 61,2 ve Bayes ağıyla %54 performans göstermiştir (Şekil 5.54). Öğrenme sadece sinir ağının tahmin kabiliyetini arttırmıştır.

Comparing Bayes_Noise with Noise-K-Means

'Partition'	1_Training	2_Testing
Correct	555 57.75%	127 54.04%
Wrong	406 42.25%	108 45.96%
Total	961	235

Comparing C5.0_Noise with Noise-K-Means

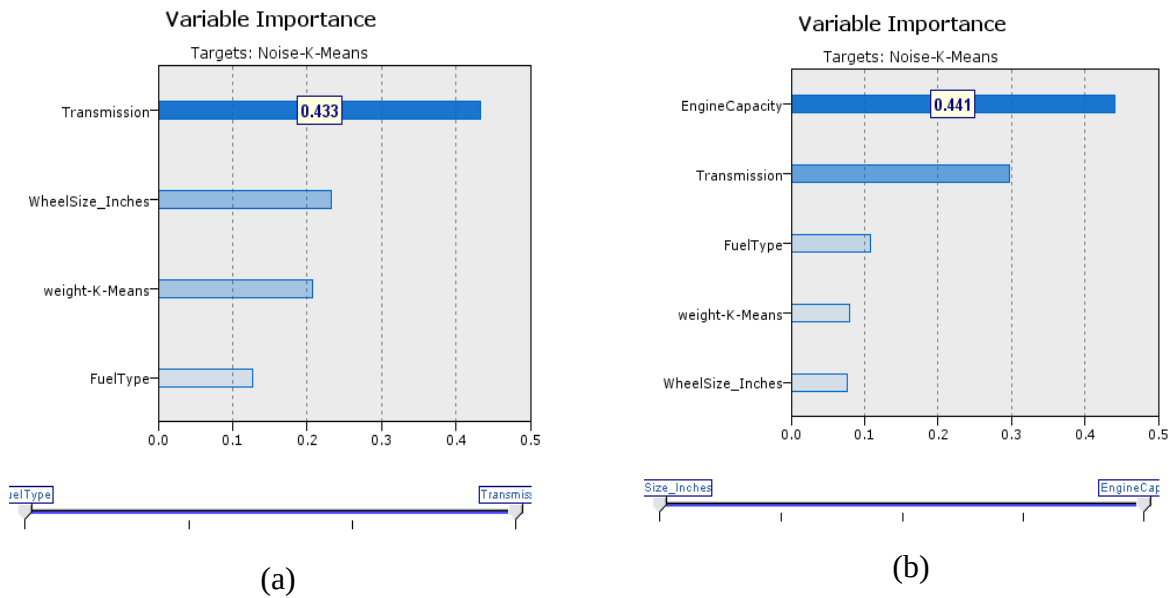
'Partition'	1_Training	2_Testing
Correct	694 72.22%	154 65.53%
Wrong	267 27.78%	81 34.47%
Total	961	235

Comparing \$N\$-Noise-K-Means with Noise-K-Means

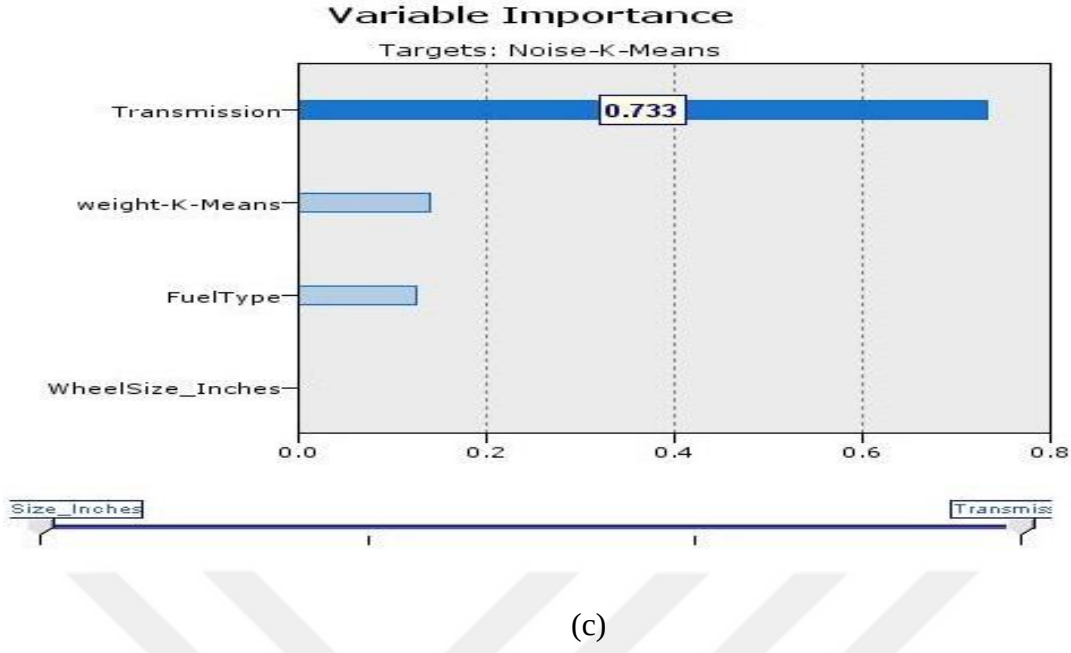
'Partition'	1_Training	2_Testing
Correct	514 53.49%	144 61.28%
Wrong	447 46.51%	91 38.72%
Total	961	235

Şekil 5.54. Gürültü seviyesi verilerine uygulanan tekniklerin analiz sonuçları

Sinir ağı ve Bayes ağı ile bulunan en önemli faktörler sırasıyla vites kutusu, araç ağırlığı, yakıt tipi ve tekerlek boyutudur (Şekil 5.55a ve Şekil 5.55b). Ağırlık dışında aynı faktörler C5.0 algoritmasıyla da önemli bulunmuştur ancak C5.0 algoritmasının verdiği sonuçlara göre motor kapasitesi bu değişkenlerden daha önemlidir (Şekil 5.55c).



Şekil 5.55. Gürültü seviyesini etkileyen (a) sinir ağı (b) Bayes ağı (c) C5.0 algoritması ile bulunan en önemli faktörler



Şekil 5.55. (devam) Gürültü seviyesini etkileyen (a) sinir ağı (b) Bayes ağı (c) C5.0 algoritması ile bulunan en önemli faktörler



6. SONUÇLAR

6.1. Faktörlerin Değerlendirilmesi

Uygulanan analizlerden sonra, Bayes ağı, sinir ağı ve C5.0 algoritmasının tahmin performansları Çizelge 6.1’de her bir faktör için karşılaştırılarak değerlendirilmiştir.

Faktörlerden yedisi (şehir içi yakıt tüketimi, kombine yakıt tüketimi, karbon monoksit emisyonu, hidrokarbon emisyonu, nitrojen oksit emisyonu, parçacık emisyonu ve gürültü seviyesi) üzerinde en etkili yöntem C5.0 algoritması olmuştur. Ancak C5.0 algoritması şehir dışı yakıt tüketimi verileri üzerinde en etkili faktörleri bulmada yanıltıcı sonuçlar vermiştir. Bunun yerine sinir ağının kullanımı hemen hemen %100 performans sağlamaktadır. Karbon dioksit salınımı üzerinde etkili olan faktörler ise Bayes ağı ile daha doğru bir biçimde tahmin edilebilmektedir.

Kullanılan tekniklerin tahmindeki genel başarısı sırasıyla Bayes ağı için %72, C5.0 algoritması için %82 ve sinir ağı için %77 olmuştur. Bu da C5.0 algoritmasının diğer iki yöntemden daha iyi sonuç verdiğini göstermektedir.

Çizelge 6.1. Sınıflandırma tekniklerinin tahmin başarısı

Fuel Consumptions/ Emissions/Noise	BN	C5.0	NN
Metric Urban Cold	72.34	87.23	82.55
Metric Extra Urban	68.06	-	99.72
Metric Combined	68.33	83.33	79.17
CO Emission	48.09	71.91	53.19
CO2 Emission	91.49	90.21	90.21
THC Emission	83.89	87.78	82.78
NO Emission	65.83	73.06	53.33
Particle Emission	95	96.11	90
Noise Level	54.04	65.53	61.28

Benzer bir karşılaştırma her bir parametre için de yapılmıştır. Çizelge 6.2’de görüldüğü gibi, yakıt tüketimi çeşitleri, emisyonlar ve gürültü seviyesini etkileyen en önemli faktörler de karşılaştırılmıştır. Buna göre sinir ağı, yakıt tüketimlerini etkileyen en önemli faktörleri

sırasıyla otomobillerin ağırlığı, vites kutusu ve yakıt tipi olarak bulmuştur. Emisyonlar üzerinde etkili faktörler ise sırasıyla vites kutusu, tekerlek boyutu ve araç ağırlığıdır. Gürültü seviyesi de benzer şekilde en fazla vites kutusu, tekerlek çapı ve ağırlıktan etkilenmektedir.

Çizelge 6.2. Sinir ağı ile bulunan parametre ağırlıkları

NN	Engine Capacity	Transmission	Wheel Size	Fuel Type	Weight	Motor Type	Series
Metric Urban	-	19.4%	12.9%	23.5%	44.2%	-	-
Metric Extra Urban	-	33.4%	24.5%	24.7%	17.4%	-	-
Metric Combined	-	13.9%	13.3%	15.2%	57.6%	-	-
CO Emission	-	42.6%	28.2%	21.7%	7.5%	-	-
CO2 Emission	-	12.5%	17.3%	10.4%	59.8%	-	-
THC Emission	-	22.4%	13.9%	39.9%	23.8%	-	-
NO Emission	-	35.8%	13.1%	28.1%	22.9%	-	-
Particle Emission	-	31.7%	14.4%	43.1%	10.8%	-	-
Noise Level	-	43.3%	23.2%	12.7%	20.2%	-	-

C5.0 algoritmasıyla elde edilen sonuçlara göre ağırlık çok önemli bir faktördür (Çizelge 6.3) ve özellikle de yakıt tüketimini büyük ölçüde etkilemektedir. Araçların serisi, kombine yakıt tüketimi için önemli bulunmuştur. Emisyonlar ise önem sırasına göre yakıt tipi, motor kapasitesi, araç ağırlığı, vites kutusu ve motor tipinden etkilenmektedir. Gerek yakıt tüketimi ve emisyonlar, gerekse gürültü seviyesi üzerinde tekerlek boyutları çok fazla etkili değildir. Sinir ağının verdiği sonuçlara benzer şekilde vites kutusu gürültü seviyesini diğer parametrelerden daha çok etkilemektedir.

Çizelge 6.3. C5.0 algoritması ile bulunan parametre ağırlıkları

C5.0	Engine Capacity	Transmission	Wheel Size	Fuel Type	Weight	Motor Type	Series
Metric Urban	11.6%	-	-	10.7%	77.5%	0.2%	-
Metric Extra Urban	-	-	-	-	100,0%	-	-
Metric Combined	2.1%	0.0%	0.3%	0.2%	0.6%	2.3%	94.5%
CO Emission	82.6%	5.2%	0.5%	5.0%	0.0%	7.5%	-
CO2 Emission	3.4%	-	-	-	96.1%	0.4%	-
THC Emission	1.4%	0.0%	-	52.4%	0.6%	45.6%	-
NO Emission	-	34.1%	0.3%	59.8%	5.8%	-	-
Particle Emission	-	34.5%	-	62.5%	2.9%	-	-
Noise Level	-	73.33%	0.0%	12.6%	14.2%	-	-

Bayes ağının verdiği sonuçlara göre motor kapasitesi ve vites kutusu tüm metrikleri etkilemektedir (Çizelge 6.4). Yakıt tüketiminin tahmininde motor kapasitesi ve ağırlık en önemli faktörlerdir. Motor kapasitesi gürültü seviyesini de diğer parametrelerden daha fazla etkilemektedir. Emisyon seviyeleri ise en fazla motor kapasitesi, yakıt tipi, araç ağırlığı ve vites kutusundan etkilenmektedir.

Çizelge 6.4. Bayes ağı ile bulunan parametre ağırlıkları

BN	Engine Capacity	Transmission	Wheel Size	Fuel Type	Weight	Motor Type	Series
Metric Urban	28.5%	7.9%	6.1%	22.4%	35.1%	-	-
Metric Extra Urban	40.9%	22.3%	13.8%	22.9%	-	-	-
Metric Combined	49.7%	28.7%	21.7%	-	-	-	-
CO Emission	72.8%	0.0%	0.0%	12.5%	14.7%	-	-
CO2 Emission	35.4%	11.4%	0.0%	20.4%	32.9%	-	-
THC Emission	45.8%	4.6%	-	31.6%	17.9%	-	-
NO Emission	46.4%	20.8%	-	22.8%	9.9%	-	-
Particle Emission	30.2%	19.6%	8.2%	19.8%	22.2%	-	-
Noise Level	44.1%	29.7%	7.6%	10.7%	7.9%	-	-

6.2. Benchmarking

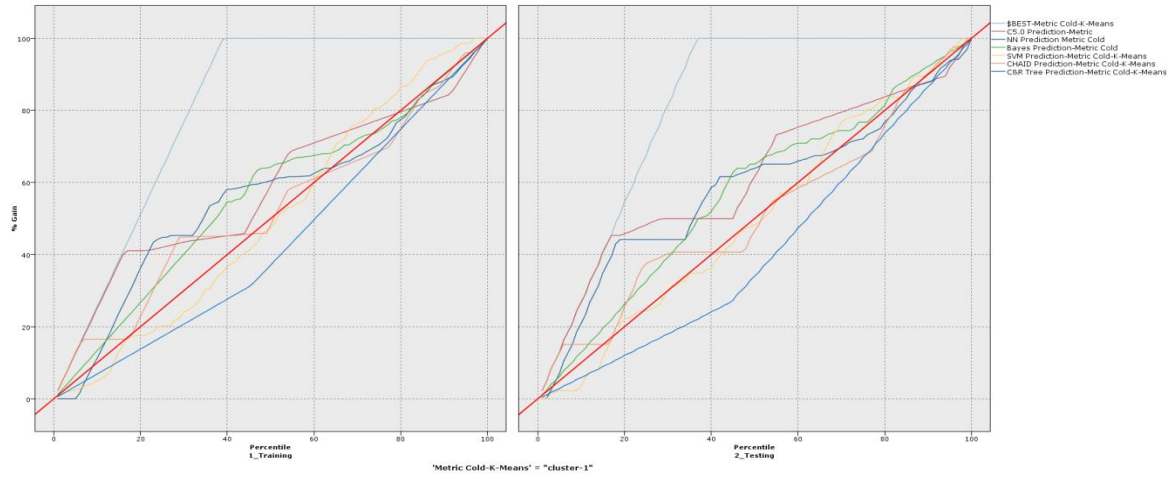
Çalışmanın kalitesini arttırmak amacıyla bu çalışmada önerilen metodolojik teknikler, daha son zamanlarda yapılmış benzer çalışmalarda tekniklerle karşılaştırılmıştır. Mevcut veri kümesine, diğer çalışmalarda farklı algoritmalar (Support Vector Machine, CHAID, CveRT) uygulanmıştır. Önerilen metodoloji ve benchmark sonuçları Şekil 6.1’de özetlenmiştir.

Comparing C5.0-Metric Cold with Metric Cold-K-Means				
'Partition'	1_Training		2_Testing	
Correct	835	86.89%	205	87.23%
Wrong	126	13.11%	30	12.77%
Total	961		235	
Comparing Bayes-Metric Colds with Metric Cold-K-Means				
'Partition'	1_Training		2_Testing	
Correct	748	77.84%	170	72.34%
Wrong	213	22.16%	65	27.66%
Total	961		235	
Comparing SVM-Metric Cold-K-Means with Metric Cold-K-Means				
'Partition'	1_Training		2_Testing	
Correct	570	59.31%	150	63.83%
Wrong	391	40.69%	85	36.17%
Total	961		235	
Comparing CHAID-Metric Cold-K-Means with Metric Cold-K-Means				
'Partition'	1_Training		2_Testing	
Correct	835	86.89%	195	82.98%
Wrong	126	13.11%	40	17.02%
Total	961		235	
Comparing C&R Tree-Metric Cold-K-Means with Metric Cold-K-Means				
'Partition'	1_Training		2_Testing	
Correct	576	59.94%	144	61.28%
Wrong	385	40.06%	91	38.72%
Total	961		235	

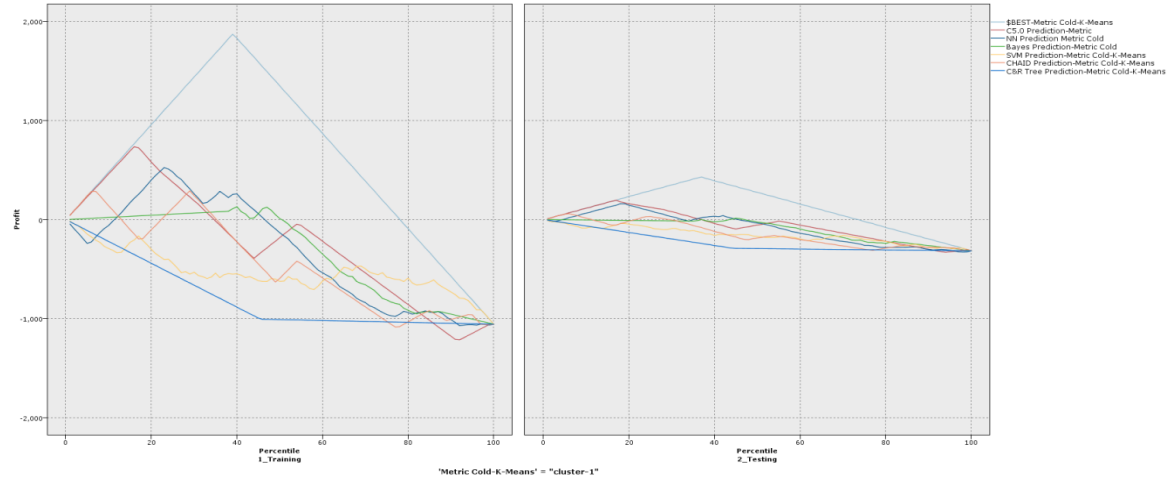
Şekil 6.1. Tekniklerin karşılaştırma sonuçları

Grafiksel gösterimlerde tüm teknikler şehir içi yakıt tüketimine uygulanarak karşılaştırılmıştır ve sonuçları Şekil 6.2’de gösterilmiştir. Önerdiğimiz metodolojide C5.0 algoritması net bir biçimde diğer yaklaşımlardan daha iyi bir performans göstermiştir.

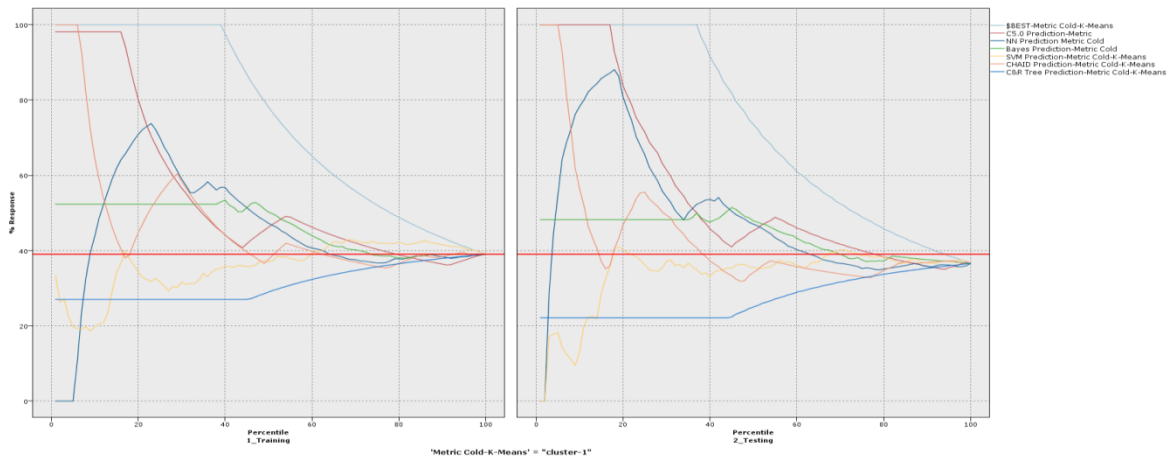
Bu çalışmada kullanılan C.5 algoritması ve Bayes ağı destek vektör makinesi (Support Vector Machine-SVM), CHAID ve C&R ağaçları ile karşılaştırılmıştır. Sonuçlar C5.0 algoritmasının tahminde %87’lik performansla diğer tekniklerden daha üstün olduğunu kanıtlamıştır. Ancak çalışmada kullanılan Bayes ağı yerine CHAID ağaçlarının tercih edilebileceği gözlemlenmiştir.



(a)



(b)



(c)

Şekil 6.2. Şehir içi yakıt tüketimi için (a)kazanç (b) kâr (c) yanıt grafiği

Şekil 6.2a’da görüldüğü gibi, şehir içi yakıt tüketimi verilerinin %39’undan itibaren %100 kazanç sağlanabilmektedir. %0-22 ve %53-68 aralıklarında C5.0 algoritması, %22-45 arasında C&R ağacı, %45-52 arasında Bayes ağı ve %52’den itibaren CHAID ideale en yakın kazanç ve faydayı vermektedir. Diğer yandan C5.0 algoritması veride %0-16 arasında ideal çözüme eşit sonuçlar vermiştir. Öğrenme en çok C5.0 algoritmasını, CHAID ve C&R ağaçlarını etkilemiş ve C5.0 algoritmasının başarısında belirgin bir artışa sebep olmuştur. Ayrıca fayda eğrileri Şekil 6.2b’deki gibi birbirlerine ve ideal çözüme yaklaşmıştır.

Şehir içi yakıt tüketimi verilerine uygulanan beş teknik için yanıt, yükseltme ve ROI grafikleri Şekil 6.2 ve EK-12’de gösterilmiştir. %0-5 aralığında CHAID ve C5.0 algoritması ideale en yakın tahminleri yapmaktadır. Verinin %18’inden itibaren C5.0 algoritmasının performansı yükselmiş ancak %18-35 ve %52-81 aralıklarında diğer tekniklerden daha iyi sonuç vermiştir. C&RT verinin %35-44 arasında, Bayes ağı ise %44-52 arasında üç grafikte de ideale en yakın tahmini vermiştir.

Önceki analizlerdeki gibi, benchmarking kısmında da tüm tekniklerin yakıt tüketimi, emisyon seviyeleri ve gürültü seviyesini etkileyen faktörleri belirlemedeki performansları karşılaştırılmıştır (Çizelge 6.5).

Çizelge 6.5. Özetlenmiş benchmarking sonuçları

Çalışma	Metot	Şehir içi yakıt tüketimi	Şehir dışı yakıt tüketimi	Kombine çevrim	CO	CO ₂	NOx	Parça-cık	THC	Gürültü Seviyesi
Önerilen	Sinir Ağı	82.5	99.7*	79.1	53.1	90.2	53.3	90	82.7	61.2
Önerilen	C5.0	87.2*	-	83.3*	71.9*	90.2	73.0	96.1*	87.7*	65.5
Önerilen	Bayes Ağı	72.3	68.6	68.3	48.0	91.4*	65.8	95	83.8	54.0
Haifeng, Wang ve Chen (2015)	SVM	63.8	73.2	61.7	53.6	74.2	80.1*	74.6	70.3	70.3
CGI GROUP INC. (2014)	CHAID	82.9	72.5	65.4	60.1	64.7	72.2	74.5	70.8	80.2*
Viswanathan (2013)	C&RT	61.2	49.9	53.2	54.2	68.4	72.4	63.4	77.2	58.5

6.3. Tartışma

Bu tezde, farklı şirketlerin otomobilleri için yakıt tüketimi ve emisyon seviyeleri baz alınarak büyük veri madenciliğinin otomotiv sektöründe nasıl uygulanabileceğini anlamak için bir strateji geliştirmeye çalışılmıştır. Bilgi boyutlarının geniş ve içerik açısından çok geniş bir yelpazeye sahip olması nedeniyle, büyük verinin daha gelişmiş veri madenciliği teknikleri uygulamayı gerektirdiğinin farkında olmak gerekmektedir. Büyük veri kümeleriyle çalışıldığında, sistem dışından basit ve düzgün istatistikleri almak artık yeterli değildir.

Tezin veri toplama aşamasındaki kısıtlar düşünüldüğünde, önerilen metodolojiyi geçerlemek açısından çeşitli veri madenciliği yaklaşımları ve teknikleri test etmek için nispeten ortalama boyutta bir veri kümesi elde edilmiştir. Temel olarak veri madenciliği zaten veriyi işlemek ve bu verideki örüntü ve trendleri tespit etmek ile ilgili olduğundan, karar verme veya yargıya varma bu çalışma için de söz konusudur. Diğer yandan, veri madenciliği sadece bilgisayar araçları, özel olarak üretilen yazılım paketleri veya veri tabanı sistemleri ile ilgili değildir. Karmaşık veri madenciliği farklı tekniklerden destek alan araçların olduğu yazılım paketlerinde tanımlanmış geçmiş olaylardan ve algoritmalarından faydalanmaktadır. Otomobillerin tasarım ve üretiminde, veri madenciliği prensiplerinin önceden de kullanıldığı düşünülebilir fakat büyük verinin ortaya çıkmasıyla daha yaygınlaşmış ve bağlantılı (connected) otomobillere zemin hazırlamıştır.

Bu tezin sonuçları, veri madenciliği araçlarının otomotive sektördeki tasarım ve diğer genişletilmiş aşamalarındaki kabiliyetlerini göstermektedir. Dahası, toplanan verilerle bağlantılı araçlar arasındaki bağlantıyla bir yol haritası oluşturmaktadır. Büyük ölçekte veri işlenmesiyle de daha karmaşık verilerin korelasyonları ve gruplar hakkında karşılaştırma yapmak ve sonuçlar elde etmek mümkündür. Veri madenciliği araçlarının kullanılması, veri depolarının ve işleme sistemleri de dâhil olmak üzere, otomotiv sektöründeki araç ve sistemlerin farklı şekillerde kullanılabilmesine imkân tanımaktadır.

6.4. Sonuçların Değerlendirilmesi

Bu tez, büyük veri madenciliğinin uygulama potansiyelinin araştırılmasında özel otomobillerin yakıt tüketimleri, emisyon seviyeleri ve gürültü seviyesinin ampirik bir analizidir. Euro 6 ve Mercedes Benz, BMW, Audi araçlarına dair teknik özellik verileri birleştirilerek elde edilen bir veri kümesi kullanılarak uygulama yapılmıştır. Betimleyici istatistikler ve veri hakkındaki bilgiler, istatistiksel araçlar ve çeşitli veri madenciliği araçlarıyla elde edilmiştir. Verilerden elde edilen bilgiler karşılaştırılmış ve (analizin sonraki aşamalarında kullanılmak üzere) kümeleme algoritmalarıyla parametreler elde edilmiştir. Tüm veri madenciliği uygulamaları CRISP DM adımları izlenerek tasarlanmış ve yürütülmüştür.

Çeşitli makine öğrenmesi araçlarının kullanılmasıyla elde edilen sonuçlar ve uygulamalar, en önemli parametrelerin uygulanan tekniğe göre değişebileceğini göstermiştir. Genel olarak C5.0 algoritması en iyi tahmini vermiştir. Her parametrenin önem derecesi, yakıt tüketimine, emisyon seviyelerine ve gürültü seviyesine katkısıyla, veri analitiği teknikleri kullanılarak değerlendirilmiştir. C5.0 algoritması faktörlerden araç ağırlığını yakıt tüketimi üzerinde, yakıt tipini emisyonlar üzerinde ve vites kutusunu gürültü seviyesi üzerinde çok önemli bulmuştur. Bayes ağı için en önemli faktörler motor kapasitesi ve araç ağırlığıdır. Bayes ağının verdiği sonuçlara göre motor kapasitesi gürültü seviyesini de diğer parametrelerden daha çok etkilemektedir. Sinir ağının verdiği sonuçlara göre, araçların ağırlığı, vites kutusu ve yakıt tipi yakıt tüketimi ve emisyon üzerinde en etkili faktörlerdir. Gürültü seviyesini sırasıyla en çok vites kutusu, tekerlek çapı ve ağırlık etkilemektedir. Benchmarking sonuçları, bu tezin metodoloji ve uygulanan teknikler açısından doğru yolda olduğunu ispatlamıştır.

Bu çalışmanın sonuçlarının kullanılması veri analitiği araçlarının farklı senaryo tiplerine dayanarak karşılaştırılabilmesine imkân tanımaktadır. Bu çalışmadan elde edilen çıkarımlar otomobillerin test tasarımlarında dikkate alınabilir. Ayrıca, araçların çevresel sertifikasyon süreci açısından araç tasarım aşamasının geliştirilmesine yardımcı olacaktır. Diğer yandan, bu tez sonuçlarının matematiksel modellerde yer alması araç rotalama problemleri ve simülasyon bazlı değerlendirmeleri geliştirecektir.

KAYNAKLAR

- Ahmed, H. (2014). *Applying Big Data Analytics for Energy Efficiency*, Yüksek Lisans Tezi, Aalto University School of Science, Finlandiya.
- Alpaydın, E. (2010), *Introduction to Machine Learning 2nd Edition*, ABD: The MIT Press Cambridge, 1-4.
- Argüden, Y. ve Erşahin, B. (2008), *Veri Madenciliği: Veriden Bilgiye, Masraftan Değere*, İstanbul: ARGE Danışmanlık A.Ş., 7-20.
- AutomotiveIT. (2013). Big Data Automotive. *Big Data Automotive Sonderedition*, 20.
- Avrupa Otomobil Üreticileri Birliği (ACEA). (2012). *Vehicle Noise - Setting the right sound levels*. Belçika: ACEA.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. ve Wirth, R. (2000). Crisp-Dm 1.0. *CRISP-DM Consortium*, 76.
- Constantinescu, Z., Marinoiu, C. ve Vladoiu, M. (2010). Driving Style Analysis Using Data Mining Techniques. *International Journal of Computers, Communications ve Control (IJCCC)*, 5(5), 654–663.
- Di Marco, F. P. ve Hanlon, R. T. (2003). *SPSS Introduction to Clementine and Data Mining*. Chicago: SPSS Inc., Chapter 11.
- Dindarloo, S. R. ve Siami-Irdemoosa, E. (2016). Determinants of fuel consumption in mining trucks. *Energy*, 112, 232–240.
- Duarte, G. O., Gonçalves, G. A., Baptista, P. C. ve Farias, T. L. (2015). Establishing bonds between vehicle certification data and real-world vehicle fuel consumption - A Vehicle Specific Power approach. *Energy Conversion and Management*, 92, 251–265.
- Edelstein, H. A. (1999). *Introduction to Data Mining and Knowledge Discovery. Data Mining and Knowledge Discovery Handbook (Vol. 2)*. ABD: Springer US, 1-15.
- Ehsani, M., Ahmadi, A., ve Fadai, D. (2016). Modeling of vehicle fuel consumption and carbon dioxide emission in road transport. *Renewable and Sustainable Energy Reviews*, 53, 1638–1648.
- Estabrooks, A. (2000). *A Combination Scheme for Inductive Learning from Imbalanced Data Sets*. Yüksek Lisans Tezi, Dalhousie University, Computer Science, Kanada.
- Fetene, G. M., Mabit, S. L., Jensen, A. F., Kaplan, S., ve Prato, C. G. (2015). *Harnessing big-data for estimating the energy consumption and driving range of electric vehicles*. 13th Network on European Communications and Transport Activities Research (NECTAR) Conference, 6623, 0–14.

- Franco, V., Sánchez, F. P., German, J. ve Mock, P. (2014). Fact sheet : europe real-world emissions from modern diesel cars. *International Council on Clean Transportation*, 45(X), 1–4.
- Gordon, K. (2007). *Principles of data management: Facilitating information sharing*. UK: British Computer Society.
- Haifeng, Z., Wang, X. H., ve Chen, X. F. (2015). Support Vector with ROC Optimization Method Based Fuel Consumption Modeling for Civil Aircraft. *Procedia Engineering*, 99, 296–303.
- Heckerman, D. (1996). A Tutorial on Learning With Bayesian Networks. *Innovations in Bayesian Networks*, 1995(November), 33–82.
- Hiziroglu, A. (2013). Soft computing applications in customer segmentation: State-of-art review and critique. *Expert Systems with Applications*, 40(16), 6491–6507.
- IBM Corp. (2011). *IBM SPSS Modeler CRISP-DM Guide*. US: IBM Corporation.
- ICCT Europe. (2013). *European Vehicle Market Statistics Pocketbook 2015/16*. International Council on Clean Transportation.
- İnternet: Anand, Vijay. Can Connected Car Save Fuel?. EE Times. 2014-03-26. URL: http://www.webcitation.org/query?url=http%3A%2F%2Fwww.eetimes.com%2Fauthor.asp%3Fdoc_id%3D1321635&date=2017-01-20. Son Erişim Tarihi: 2017-01-20.
- İnternet: Band, I. Car fuel data, CO2 and vehicle tax. URL: <http://www.webcitation.org/query?url=http%3A%2F%2Fcarfueldata.direct.gov.uk%2F&date=2017-01-20>. Son Erişim Tarihi: 2017-01-20.
- İnternet: Brünglinghaus, Christiane. "Big Brother" im Auto. 2014-02-05. URL: <http://www.webcitation.org/query?url=https%3A%2F%2Fwww.springerprofessional.de%2Fautomobilelektronik---software%2Fcar-to-x%2Fbig-brother-im-auto%2F6561256%3Fredirect%3D1&date=2017-01-20>. Son Erişim Tarihi: 2017-01-17.
- İnternet: CGI Group Inc. Modeling the Relation Between Driving Behavior and Fuel Consumption. 2017-01-17. URL: http://www.webcitation.org/query?url=https%3A%2F%2Fwww.cgi.com%2Fsites%2Fdefault%2Ffiles%2Fwhite-papers%2Fdriving_behavior_and_fuel_consumption.pdf&date=2017-01-20. Son Erişim Tarihi: 2017-01-20.
- İnternet: CSC. Big Data in the Automotive Industry. 2017-01-18. URL: http://www.webcitation.org/query?url=http%3A%2F%2Fwww.csc.com%2Fauto%2Finsights%2F103241-big_data_in_the_automotive_industry&date=2017-01-20. Son Erişim Tarihi: 2017-01-20.

İnternet: Does Your Company Have a Data Science Strategy?. Harvard Business Review. 2017-01-20. URL: <http://www.webcitation.org/query?url=https%3A%2F%2Fhbr.org%2Fsponsored%2F2016%2F08%2Fdoes-your-company-have-a-data-science-strategy.html+&date=2017-01-20>. Son Erişim Tarihi: 2017-01-20.

İnternet: Fowler, G. A. (August, 2015). Connecting Your Car to Your Smartphone Can Make Auto Data Work for You. *The Wall Street Journal*, 1–6. URL: <http://www.webcitation.org/query?url=http%3A%2F%2Fwww.wsj.com%2Farticles%2Fdevices-get-cars-to-reveal-secrets-about-the-engineand-how-you-drive-1440528524.html+&date=2017-01-20>. Son Erişim Tarihi: 2017-01-20.

İnternet: Fujitsu. Fujitsu develops highaccuracy fuel efficiency estimates through a ship's operational data. Phys.org. 2016-05-10. URL: <http://www.webcitation.org/query?url=http%3A%2F%2Fphys.org%2Fnews%2F201605fujitsuhighaccuracyfuelefficiencyship.html&date=2017-01-20>. Son Erişim Tarihi: 2017-01-20.

İnternet: Gottlob, G. Extracting data from the web. (2016) 2017-01-18. URL: <http://www.webcitation.org/query?url=http%3A%2F%2Fwww.extractingdata.com&date=2017-01-20>. Son Erişim Tarihi: 2017-01-20.

İnternet: IBM Corp. IBM Analytics. 2017-01-18. URL: <http://www.webcitation.org/query?url=http%3A%2F%2Fwww.ibm.com%2Fanalytics%2Fus%2Fen%2Ftechnology%2Fspss&date=2017-01-20>. Son Erişim Tarihi: 2017-01-20.

İnternet: Kielman, C., ve Dettling, J. (March, 2015).. *Computer Woche*, 4. URL: <http://www.webcitation.org/query?url=http%3A%2F%2Fwww.computerwoche.de%2Fa%2Fmit-big-data-zum-connected-car-der-zukunft%2C3094413%23.html+&date=2017-01-20>. Son Erişim Tarihi: 2017-01-20.

İnternet: LeBoeuf, Kelly. The 5 Vs of Big Data: Predictions for 2016. Excelacom. 2017-01-18. URL: <http://www.webcitation.org/query?url=http%3A%2F%2Fwww.excelacom.com%2Fresources%2Fblog%2Fthe-5-vs-of-big-data-predictions-for-2016&date=2017-01-20>. Son Erişim Tarihi: 2017-01-20.

İnternet: Leveraging “Big Data” to Maximize Fuel Economy. Fleet Owner. 2017-01-20. URL: <http://www.webcitation.org/query?url=http%3A%2F%2Ffleetowner.com%2Fleveraging-big-data-maximize-fuel-economy&date=2017-01-20>. Son Erişim Tarihi: 2017-01-20.

İnternet: Mercedes-Benz. (2016). Mercedes me connect. URL: <http://www.webcitation.org/query?url=https%3A%2F%2Fwww.mercedes-benz.com%2Fen%2Fmercedes-me%2Fconnectivity.html&date=2017-01-20>. Son Erişim Tarihi: 2017-01-20.

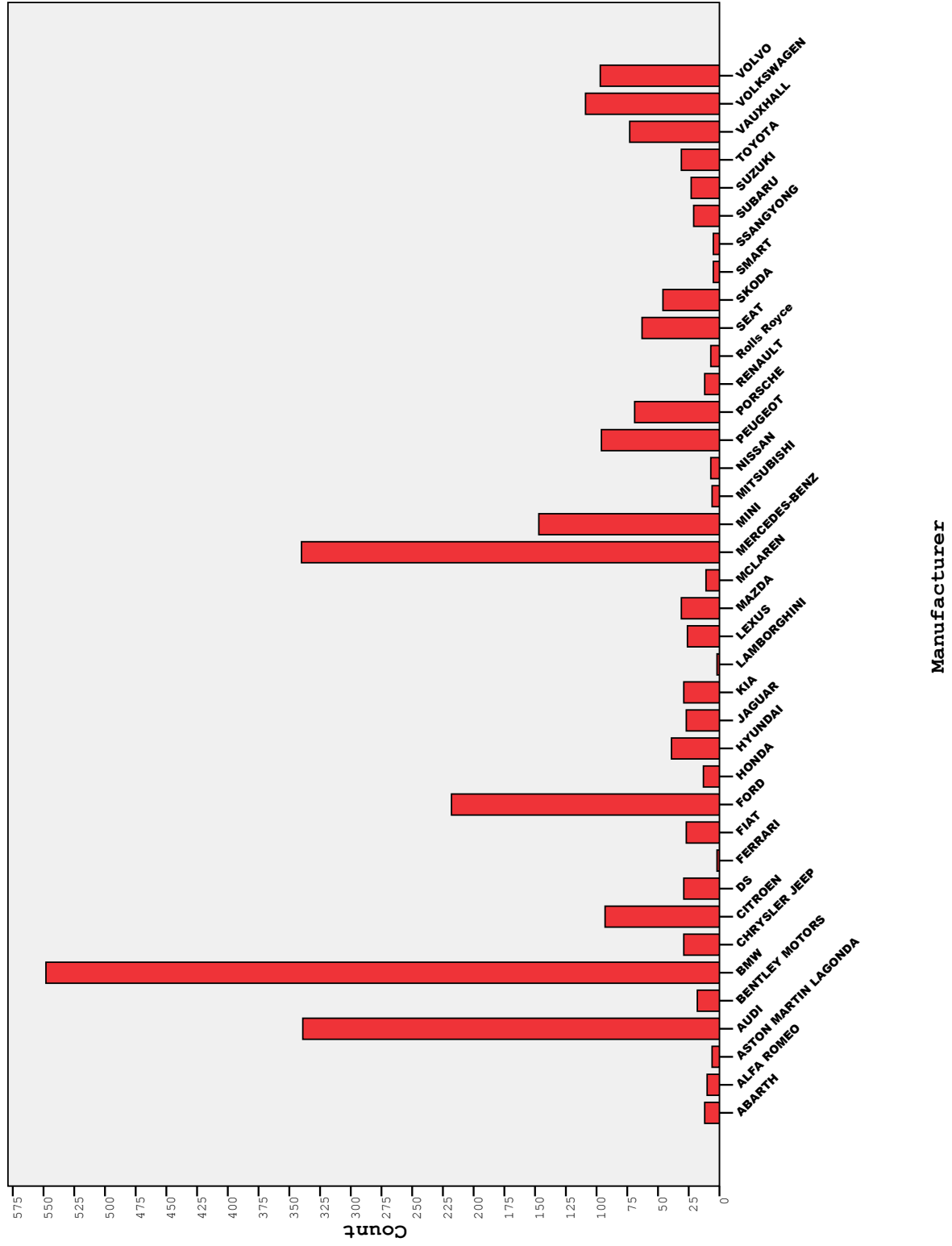
- Internet: O'Connor, Chris. How the Internet of Things is improving performance on the Formula One power unit operation. 2016-03-17. URL: <http://www.webcitation.org/query?url=http%3A%2F%2Fwww.ibmbigdatahub.com%2Fblog%2Fhow-internet-things-improving-performance-formula-one-circuit.html&date=2017-01-20>. Son Eriřim Tarihi: 2017-01-20.
- Internet: Rijmenam, Mark van. Self-driving Cars Will Create 2 Petabytes Of Data, What Are The Big Data Opportunities For The Car Industry? . 2017-01-18. URL: <http://www.webcitation.org/query?url=https%3A%2F%2Fdatafloq.com%2Fread%2Fself-driving-cars-create-2-petabytes-data-annually%2F172.html&date=2017-01-20>. Son Eriřim Tarihi: 2017-01-20.
- Internet: Towers, C. Ford's Use Of Big Data. The Innovation Enterprise. 2017-01-18. URL: <http://www.webcitation.org/query?url=https%3A%2F%2Fchannels.theinnovationenterprise.com%2Farticles%2Fford-s-use-of-big-data&date=2017-01-20>. Son Eriřim Tarihi: 2017-01-20.
- Internet: XL Hybrids' system measures MPG, fuel savings and CO2 reduction. Fleet Owner. 2017-01-20. URL: <http://www.webcitation.org/query?url=http%3A%2F%2Ffleetowner.com%2Fpowertrain%2Fxl-hybrids-system-measures-mpg-fuel-savings-and-co2-reduction&date=2017-01-20>. Son Eriřim Tarihi: 2017-01-20.
- Jin, X., Wah, B. W., Cheng, X. ve Wang, Y. (2015). Significance and Challenges of Big Data Research. *Big Data Research*, 2(2), 59–64.
- Johanson, M., Belenki, S., Jalminger, J., Fant, M. ve Gjertz, M. (2015). *Big automotive data: Leveraging large volumes of data for knowledge-driven product development*. Proceedings - 2014 IEEE International Conference on Big Data, IEEE Big Data 2014, 736–741.
- Johnsen, S., ve Felldin, S. (2016). *Improving knowledge of truck fuel consumption using data analysis*, Yüksek Lisans Tezi, Linköping Üniversitesi, İsveç.
- King, T. (2007). The automotive industry. *European Voice*, 1-16.
- Kumar, S. ve Ravishankar, K. (2015). A Study on Real Time Big Data Analytics, *International Journal of Innovative Research in Computer and Communication Engineering*, 3(8), 7553-7558.
- Li, L., Wang, X. ve Song, J. (2016). Fuel consumption optimization for smart hybrid electric vehicle during a car-following process, *Mechanical Systems and Signal Processing*, 87(B15), 17–29.
- Li, X. ve Sun, J. (2014). Effect of interactions between vehicles and pedestrians on fuel consumption and emissions. *Physica A: Statistical Mechanics and Its Applications*, 416, 661–675.

- Lynette, C., Christopher, E., Anup, B. ve John, H. (2007). Factor of Two: Halving the Fuel Consumption of New U.S. Automobiles by 2035, *Publication No. LFEE 2007-04 RP, Laboratory for Energy and Environment Massachusetts Institute of Technology*, 1–51.
- Martin, N. P. D., Bishop, J. D. K., Choudhary, R., ve Boies, A. M. (2014). Can UK passenger vehicles be designed to meet 2020 emissions targets? A novel methodology to forecast fuel consumption with uncertainty analysis. *Applied Energy*, 157, 929–939.
- Mercedes Benz. Technical specification. 2017-01-18. URL:http://www.mercedes-benz.co.uk/content/unitedkingdom/mpc/mpc_unitedkingdom_website/en/home_mpc/passengercars/home/new_cars/models/e-class/s213.html. Son Erişim Tarihi: 2017-01-18.
- MHP (2014). *MHP AMG Projekt*. BITKOM Big Data Summit 2014, Almanya.
- Mooi, E. ve Sarstedt, M. (2011). A Concise Guide to Market Research The Process, Data, and Methods Using IBM SPSS Statistics Introduction to Market Research. *Springer-Verlag Berlin, Heidelberg*, 25-284.
- Nakhaeizadeh, G. ve Schnabl, A. (1997). Development of Multi-Criteria Metrics for Evaluation of Data Mining Algorithms. *International Conference on Knowledge Discovery in Database ve Data Mining*, 37–42.
- Neelamegam, S. ve Ramaraj, E. (2013). Classification algorithm in data mining: An Overview. *International Journal of P2P Network Trends and Technology (IJPTT)*, 4(8), 369–374.
- Offices, V. (2016). *The Fuel Consumption Testing Scheme*, Birleşik Krallık: Department for Transport, Vehicle Certification Agency.
- Pakhira, M., Bandyopadhyay, S. ve Maulik, U. (2004). Validity index for crisp and fuzzy clusters. *Pattern Recognition*, 37(3), 487–501.
- Rokach, L. ve Maimon, O. (2010). *The Data Mining and Knowledge Discovery Handbook*, Chapter 15— Clustering methods, ABD: Springer US, 321-352.
- Sarvari, P. A., Üstündag, A. ve Takçı, H. (2016). Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis. *Kybernetes*, 45(7), 1129–1157.
- Siarni-Irdemoosa, E. ve Dindarloo, S. R. (2015). Prediction of fuel consumption of mining dump trucks: A neural networks approach. *Applied Energy*, 151, 77–84.
- SPSS Inc. (2007). *Clementine 11.1 User's Guide*. ABD: SPSS Inc.
- The UK Department for Transport (2016). *Vehicle Emissions Testing Programme - Moving Britain Ahead (April)*. Birleşik Krallık: The UK Department for Transport.
- Tobergte, D. ve Curtis, S. (2013). *Business Intelligence. Journal of Chemical Information and Modeling* (Vol. 53). Birleşik Krallık: Cambridge University Press.

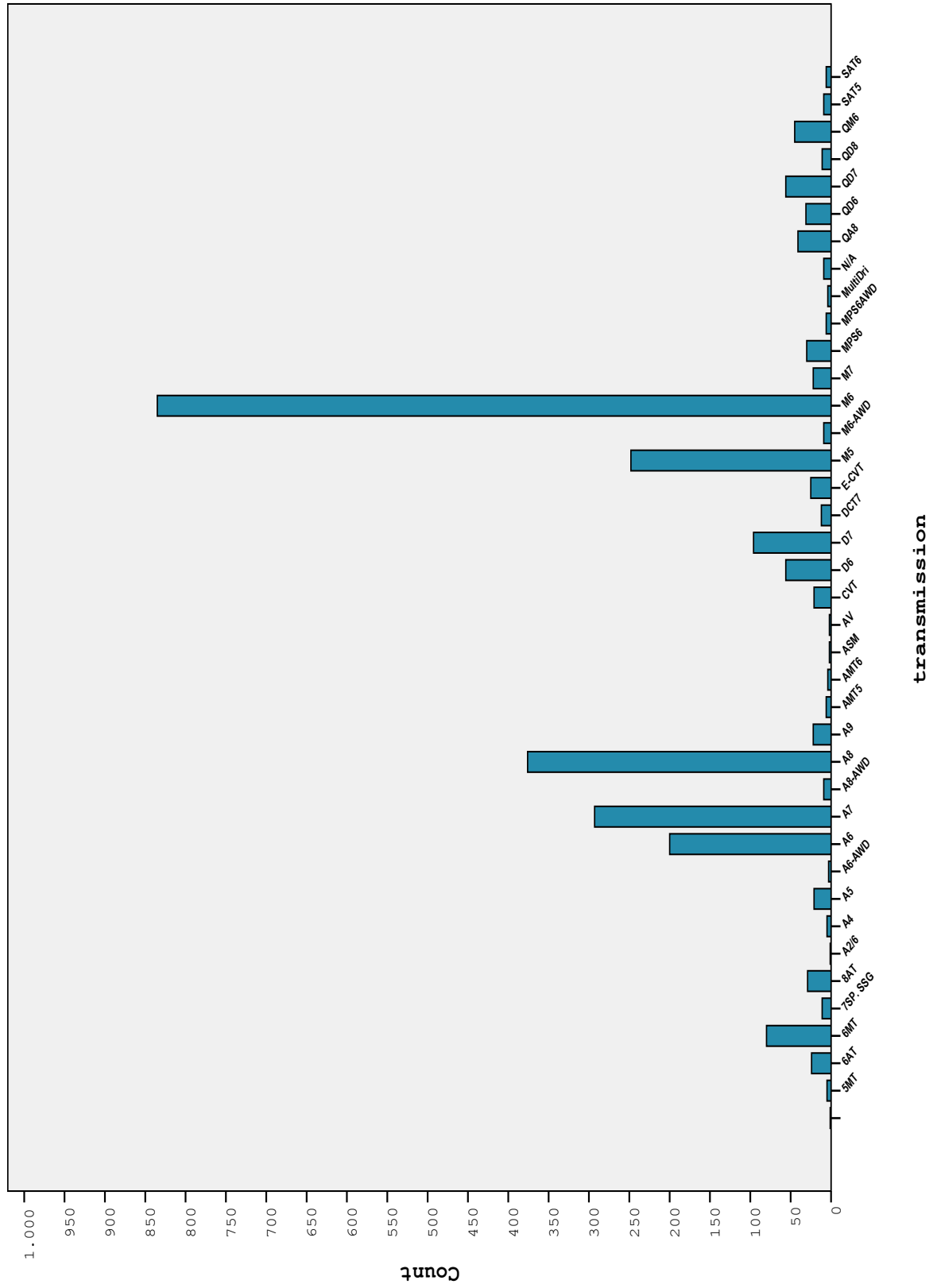
- T-Systems. (2013). *Connected Car: Traffic and Diagnostics. Connecting Vehicles-Connecting customers. Big data in the Automotive industry*, Almany: T-Systems.
- Viswanathan, A. (2013). *Data driven analysis of usage and driving parameters that affect fuel consumption of heavy vehicles*. Yüksek Lisans Tezi, Linköping University. İsveç.
- Wang, H., Zhang, X. ve Ouyang, M. (2015). Energy consumption of electric vehicles based on real-world driving patterns : A case study of Beijing. *Applied Energy*, 157, 710–719.
- Wang, X., Chow, J., Kohl, S., Percy, K., Legge, A., ve Watson, J. (2015). Real-world emission factors for Caterpillar 797B heavy haulers during mining operations. *Particuology*, 28, 22–30.
- Zamboni, G., André, M., Roveda, A. ve Capobianco, M. (2015). Experimental evaluation of Heavy Duty Vehicle speed patterns in urban and port areas and estimation of their fuel consumption and exhaust emissions. *Transportation Research Part D: Transport and Environment*, 35, 1–10.
- Zeng, W., Miwa, T. ve Morikawa, T. (2016). Prediction of vehicle CO2 emission and its application to eco-routing navigation. *Transportation Research Part C: Emerging Technologies*, 68, 194–214.
- Zhang, J., Zhao, Y., Xue, W. ve Li, J. (2015). Vehicle routing problem with fuel consumption and carbon emission. *International Journal of Production Economics*, 170, 234–242.
- Zhang, S., Wu, Y., Liu, H., Huang, R., Un, P., Zhou, Y., Hao, J. (2014). Real-world fuel consumption and CO2 (carbon dioxide) emissions by driving conditions for light-duty passenger vehicles in China. *Energy*, 69, 247–257.
- Zhang, S., Wu, Y., Un, P., Fu, L. ve Hao, J. (2016). Modeling real-world fuel consumption and carbon dioxide emissions with high resolution for light-duty passenger vehicles in a traffic populated city. *Energy*, 113, 461–471.



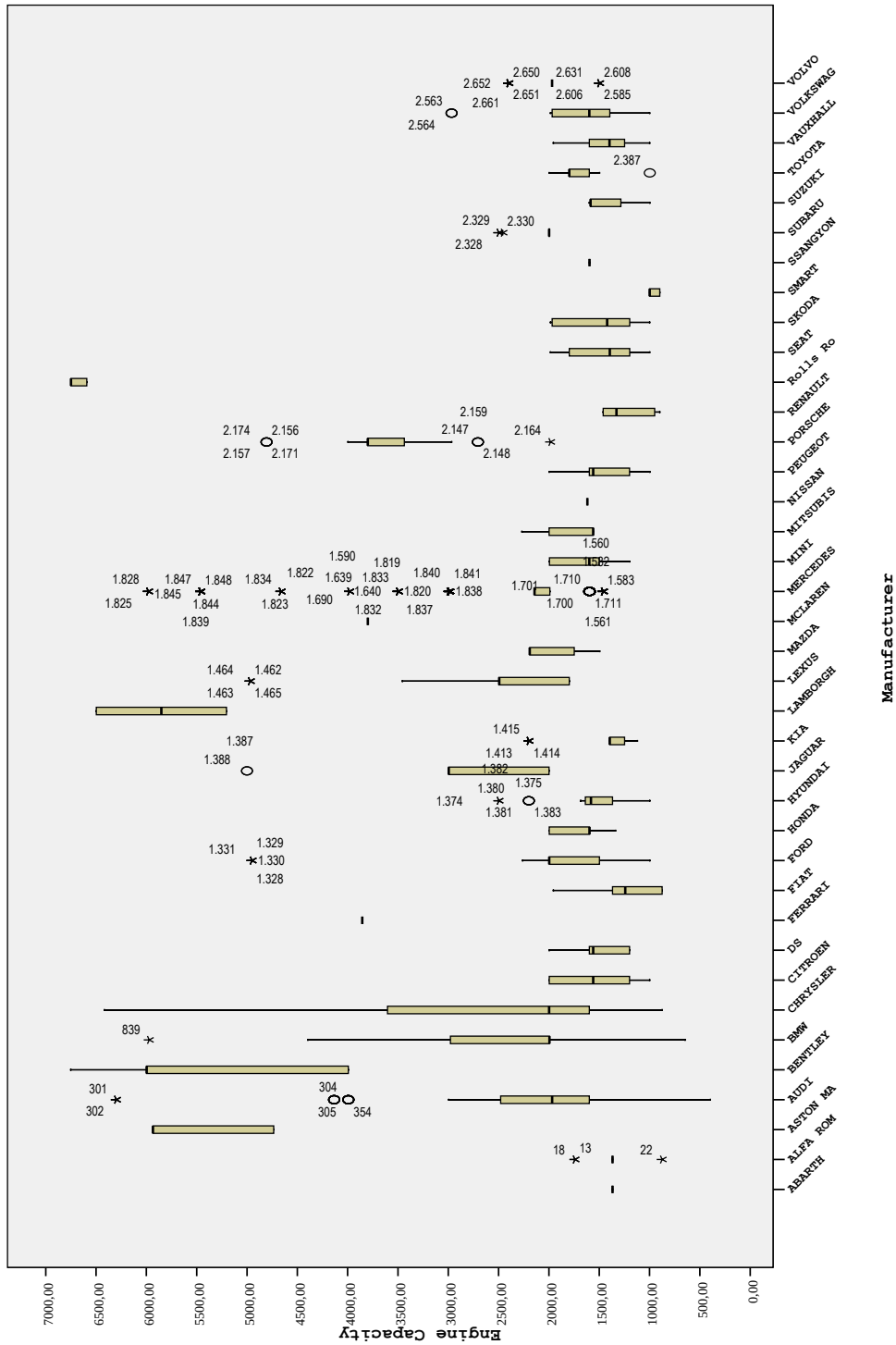
EK-1. Üreticilerin dağılımını gösteren histogram



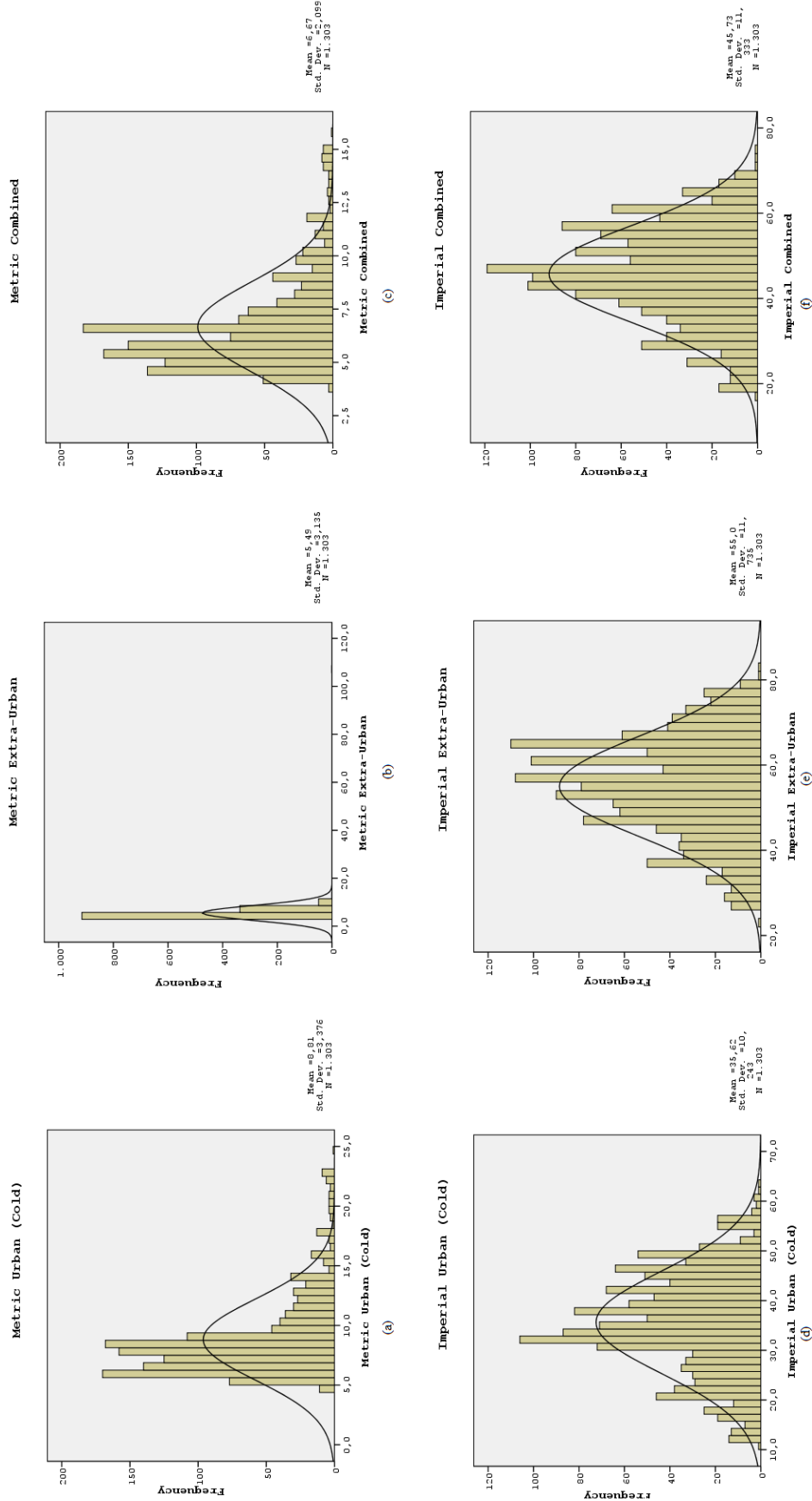
EK-2 Şanzıman tiplerinin dağılımı



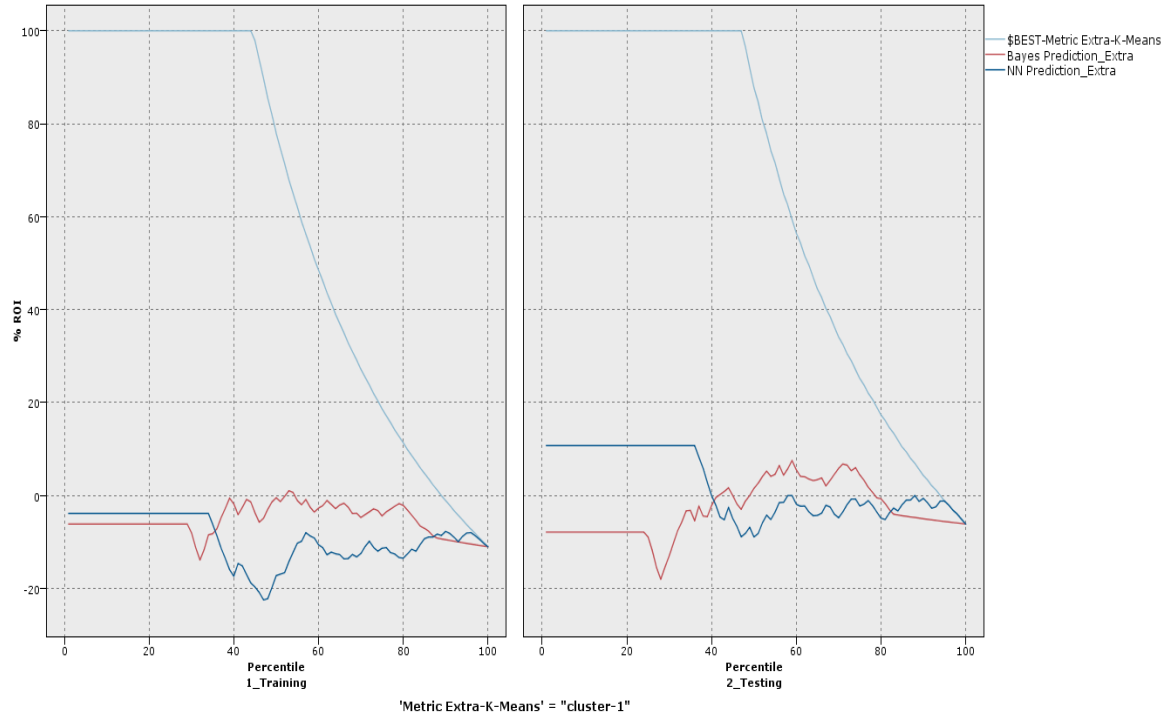
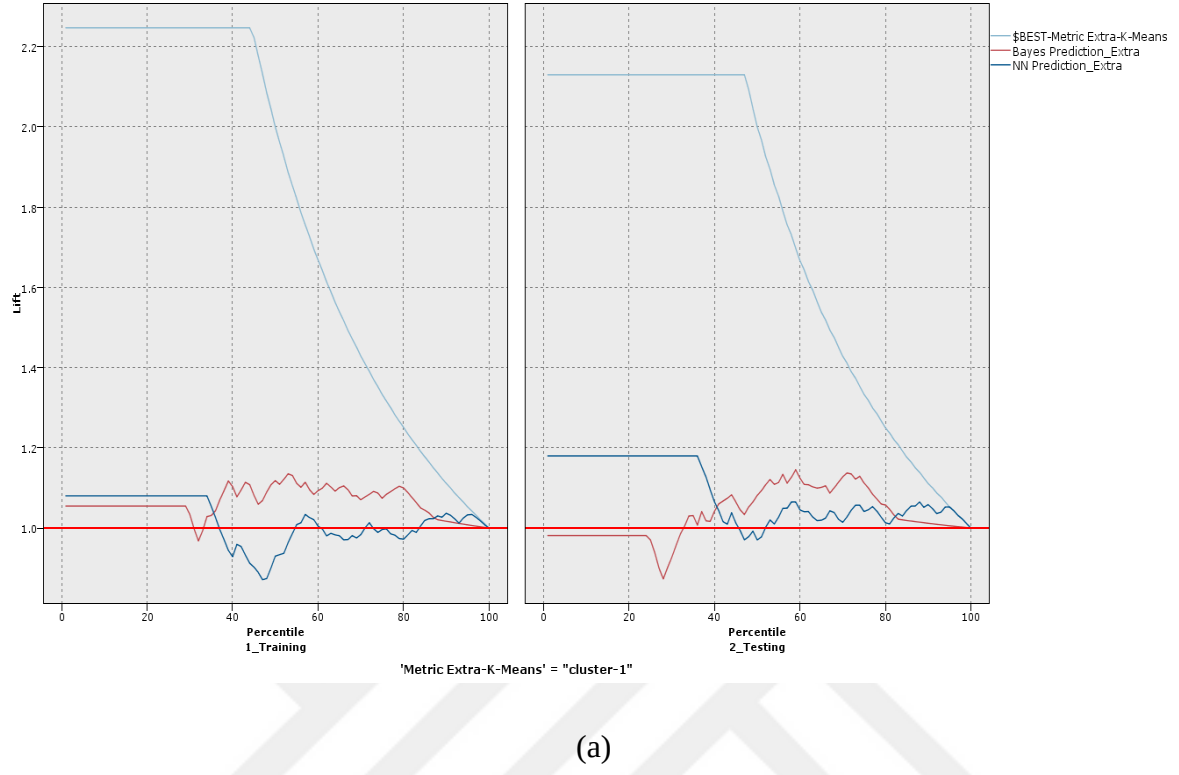
EK-3. Her bir üreticinin motor kapasitesini gösteren kutu diyagramları



EK-4. Test çevrimlerine göre yakıt tüketimi dağılımları

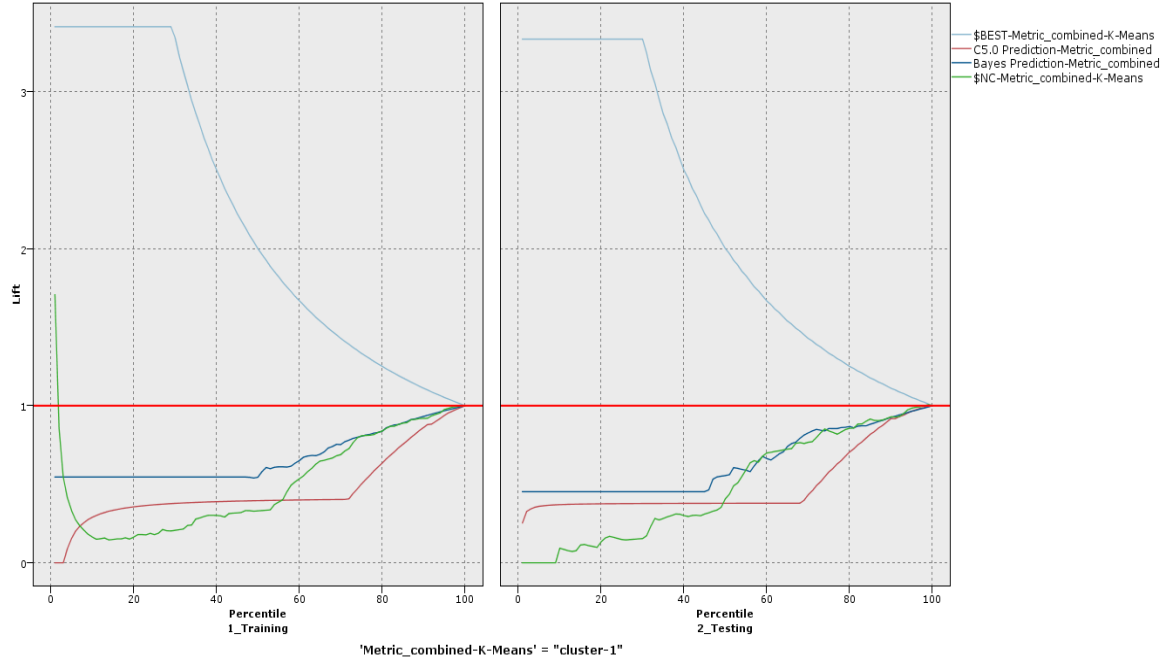


EK-5. Şehir dışı yakıt tüketimi (a) yükseltme (b) ROI grafiği

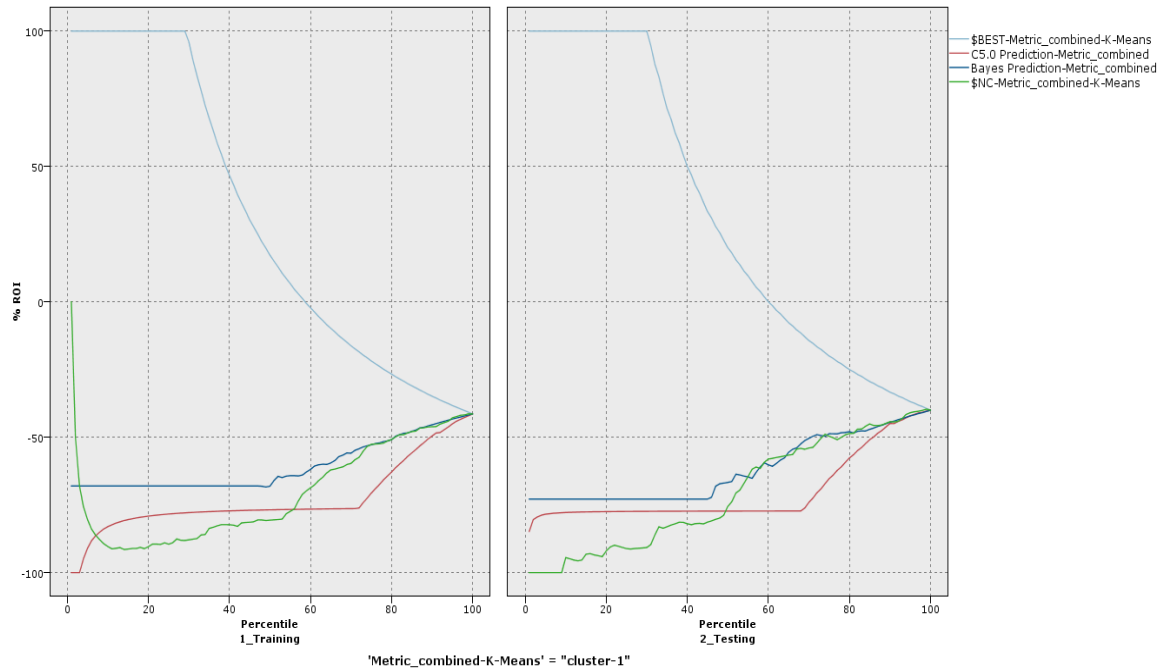


(b)

EK-6. Kombine yakıt tüketimi için (a) yükseltme (b) ROI grafiği

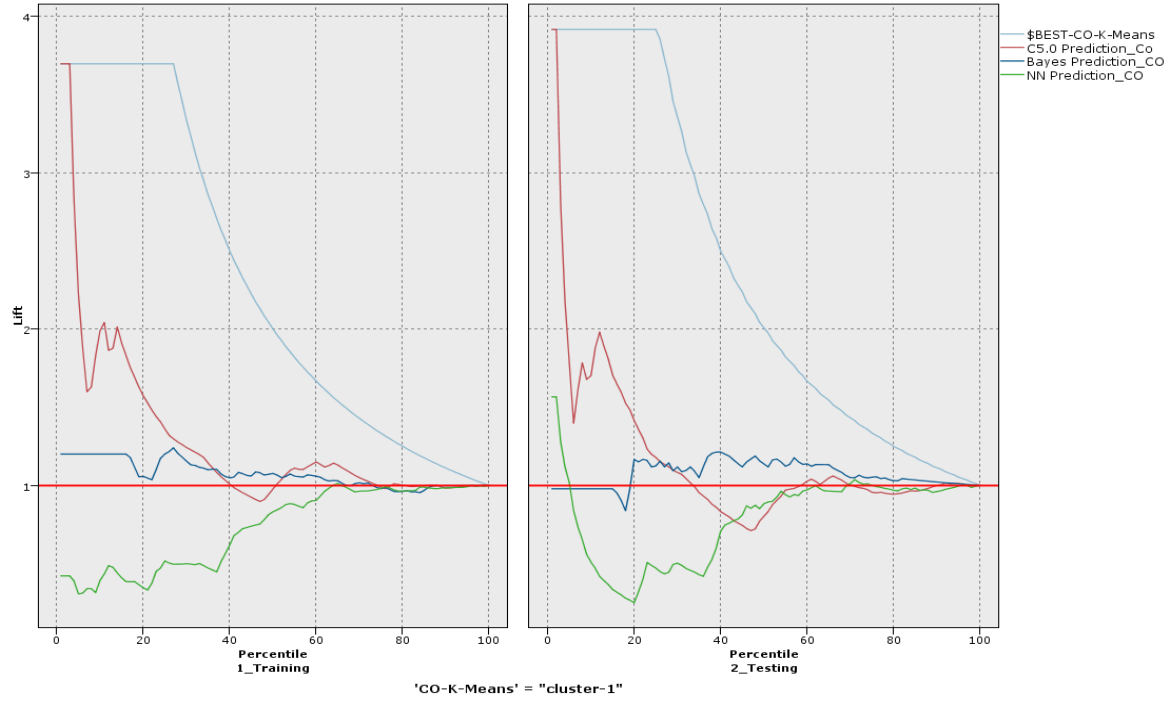


(a)

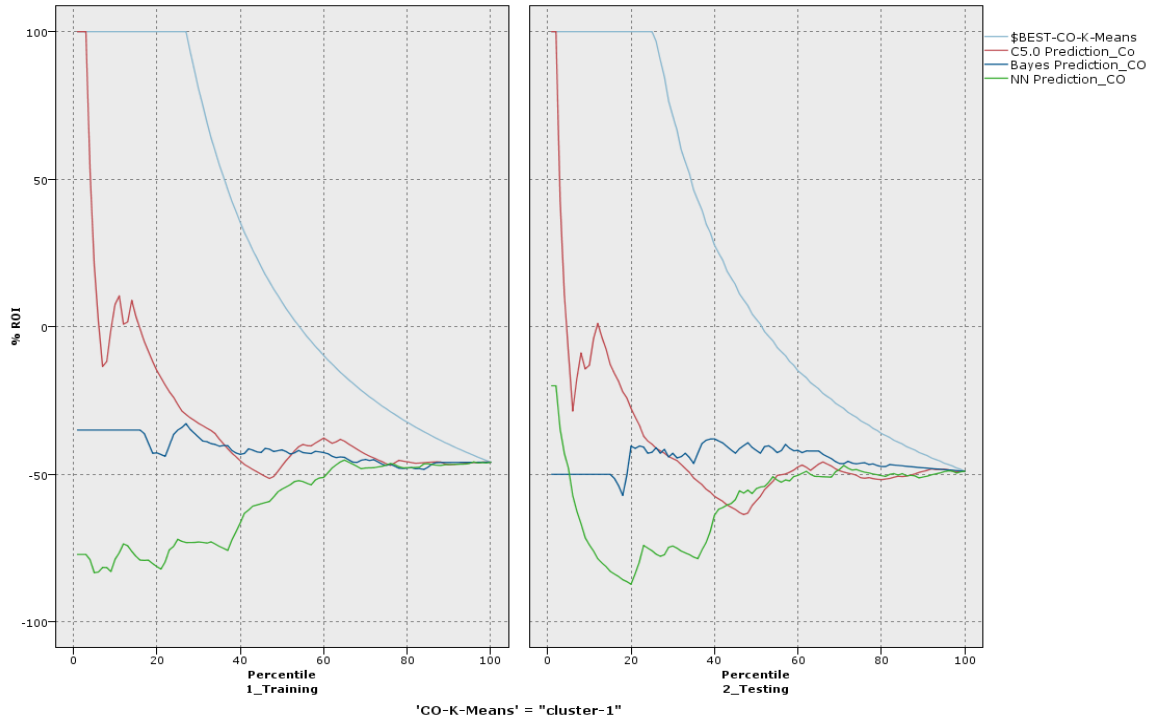


(b)

EK-7. Karbon monoksit emisyonu için (a) yükseltme (b) ROI grafiği

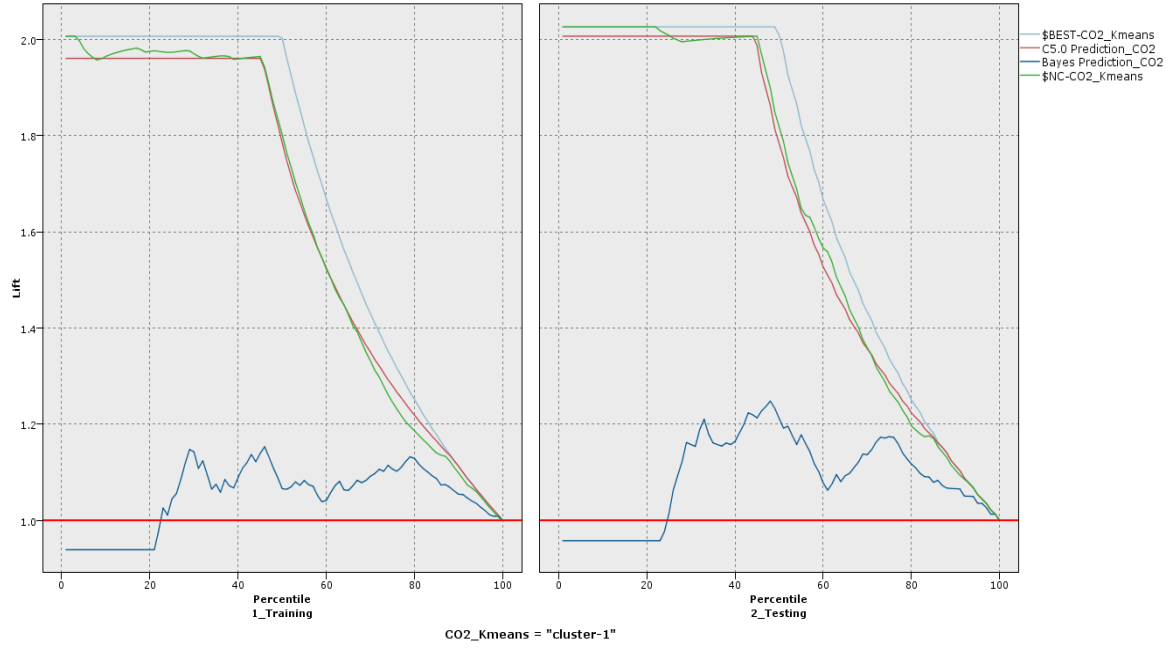


(a)

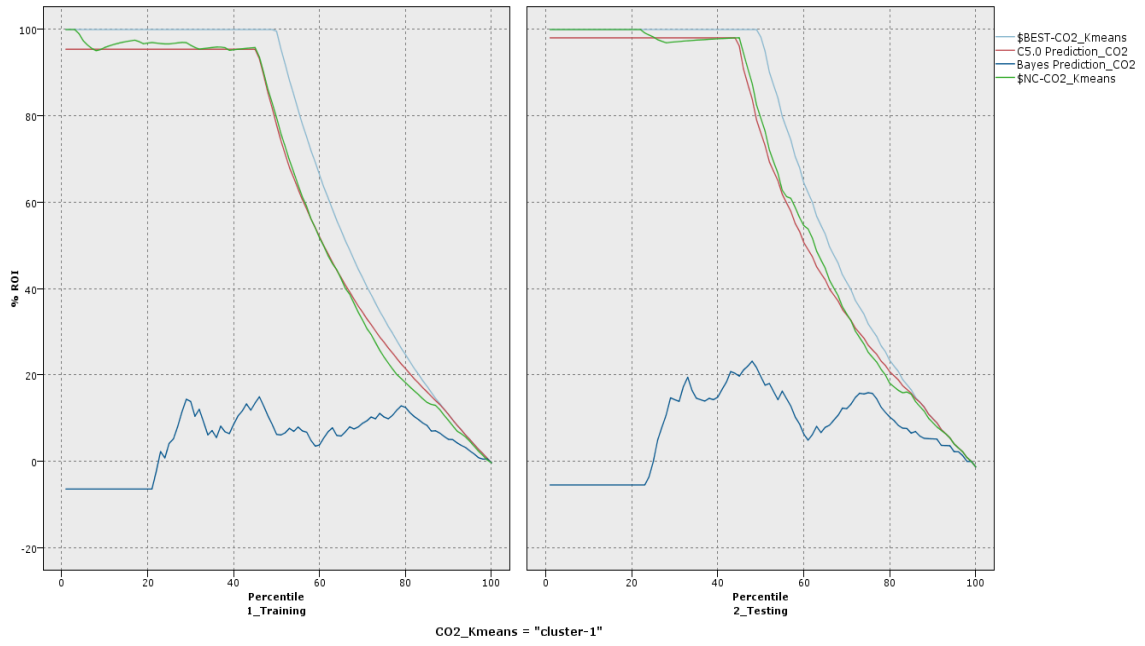


(b)

EK-8 Karbon dioksit emisyonu için (a) yükseltme (b) ROI grafiği

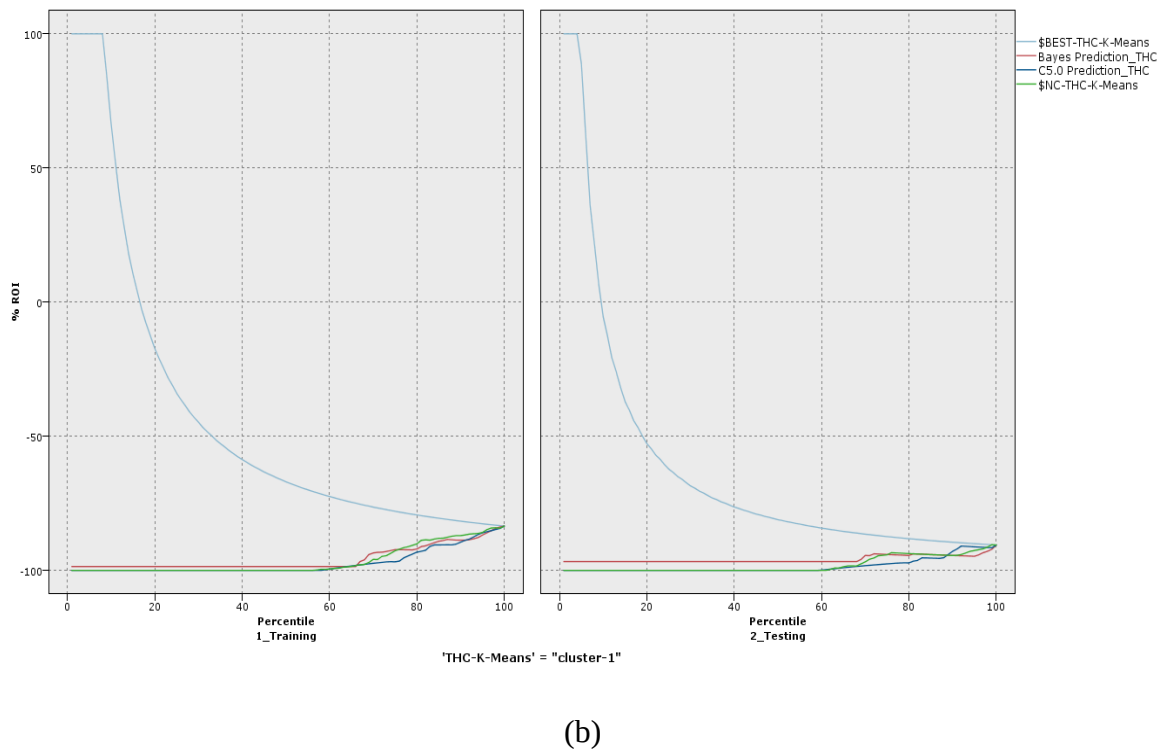
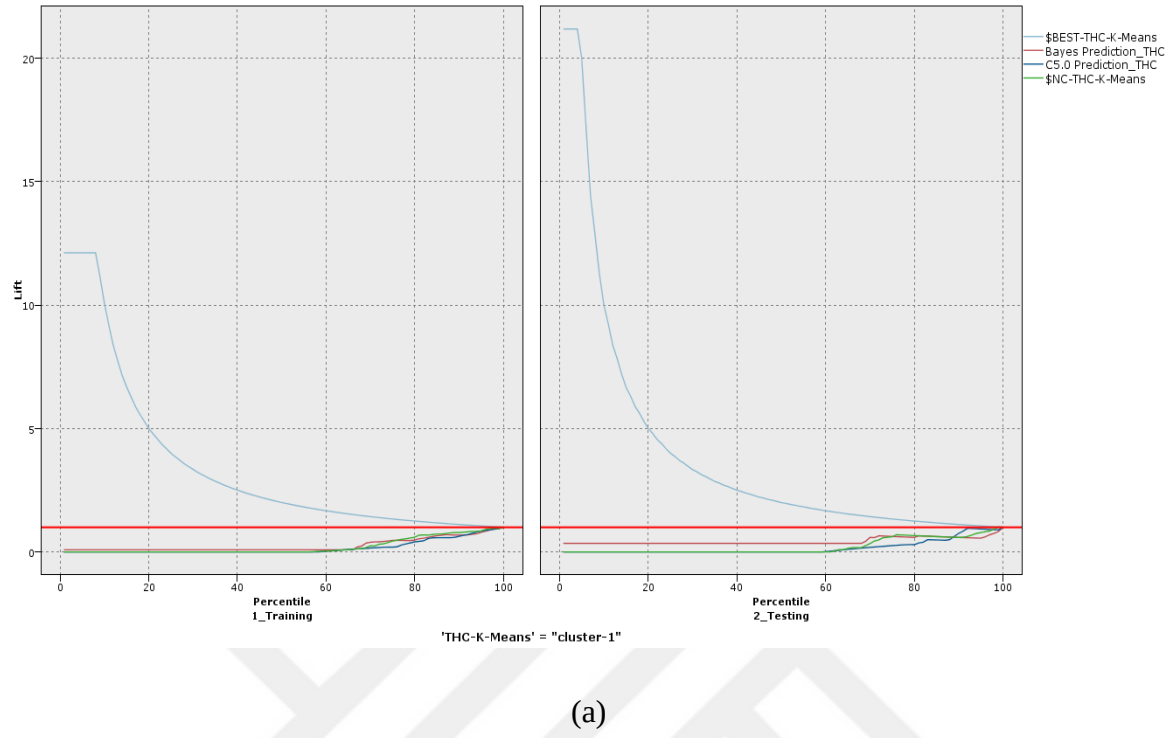


(a)

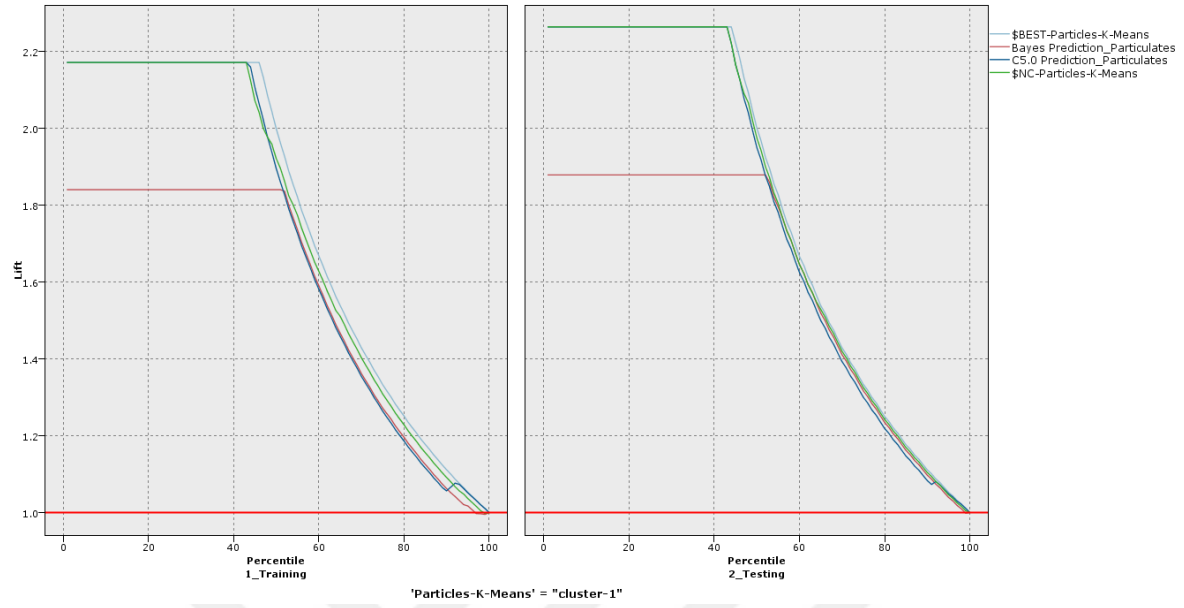


(b)

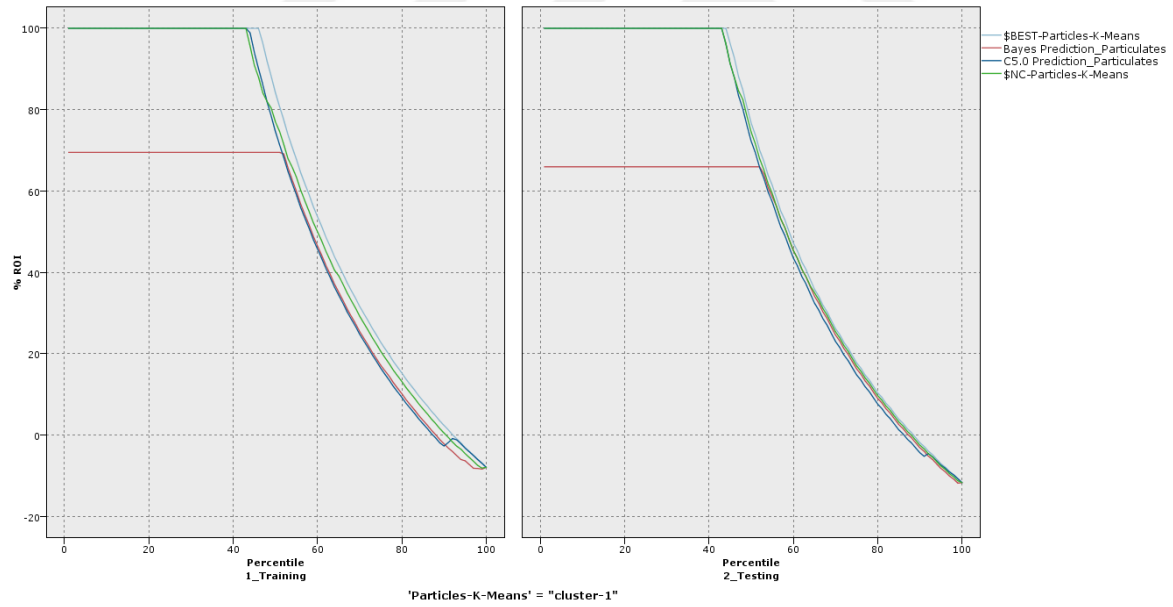
EK-9 Hidrokarbon emisyonu için (a) yükseltme (b) ROI grafiği



EK-10 Paracık emisyonu iin (a) yükseltme (b) ROI grafiđi

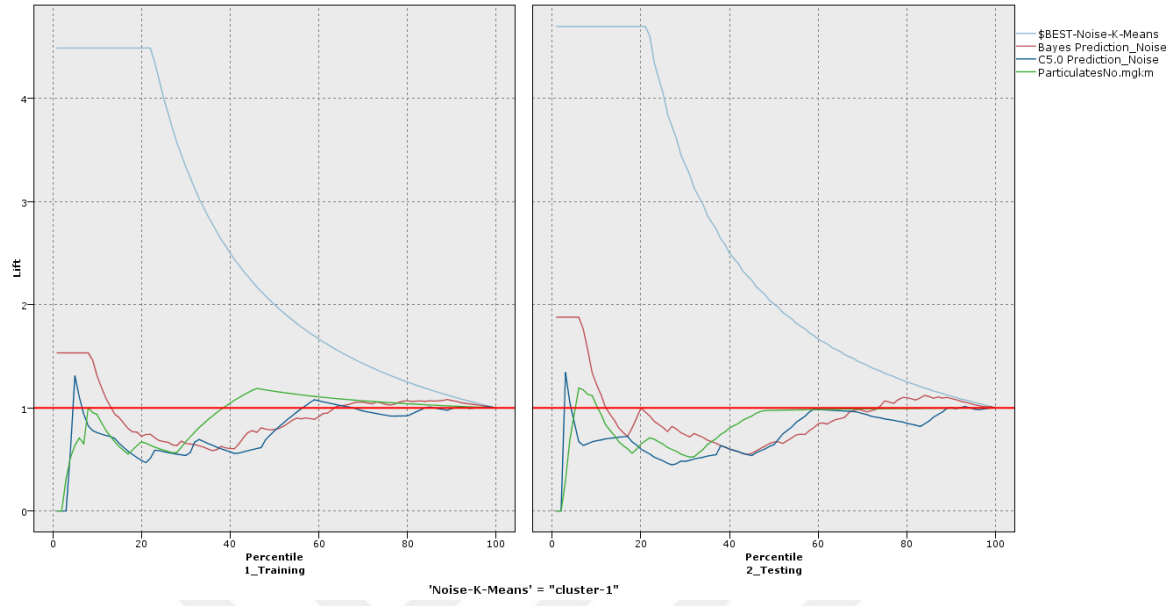


(a)

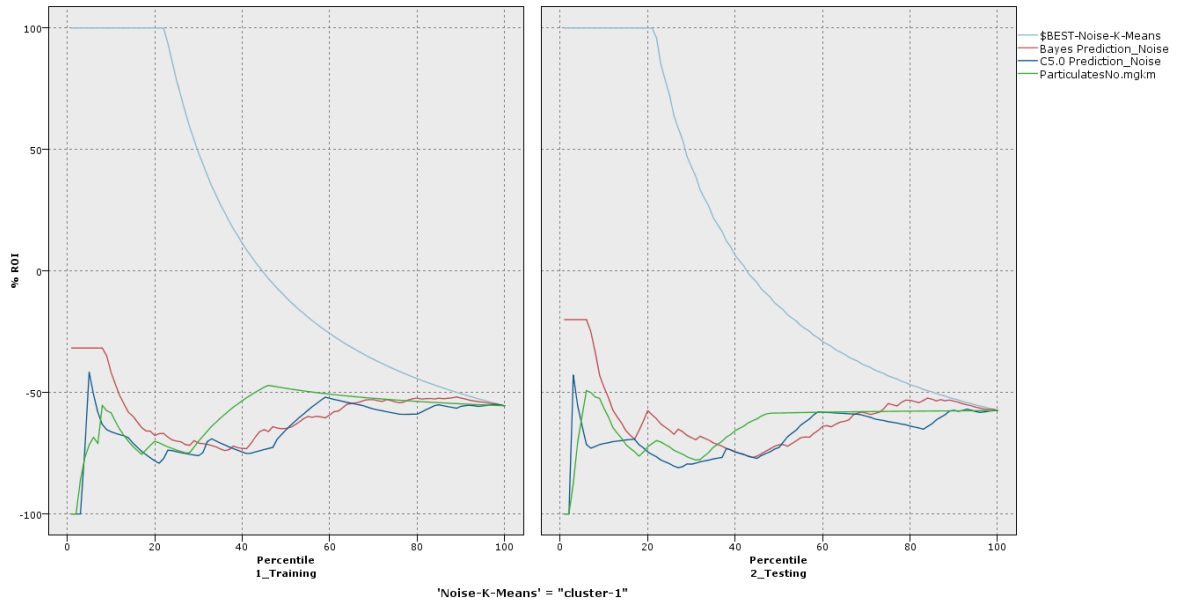


(b)

EK-11 Gürültü seviyesi için (a) yükseltme (b) ROI grafiği

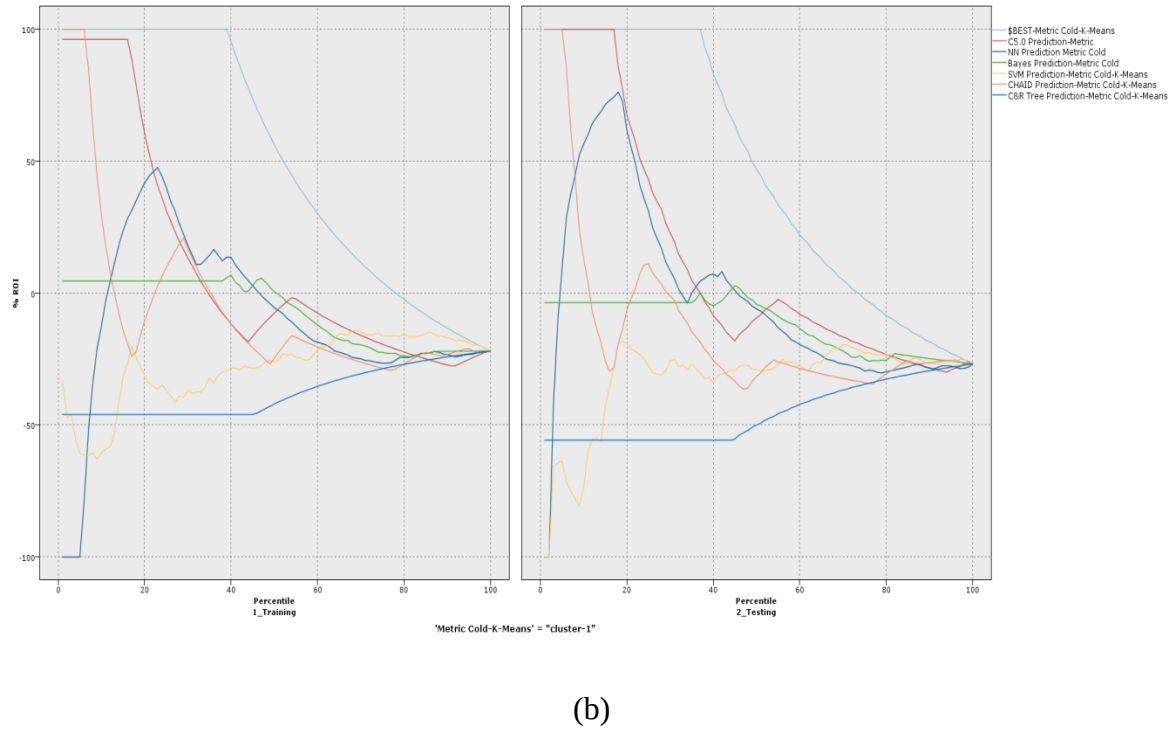
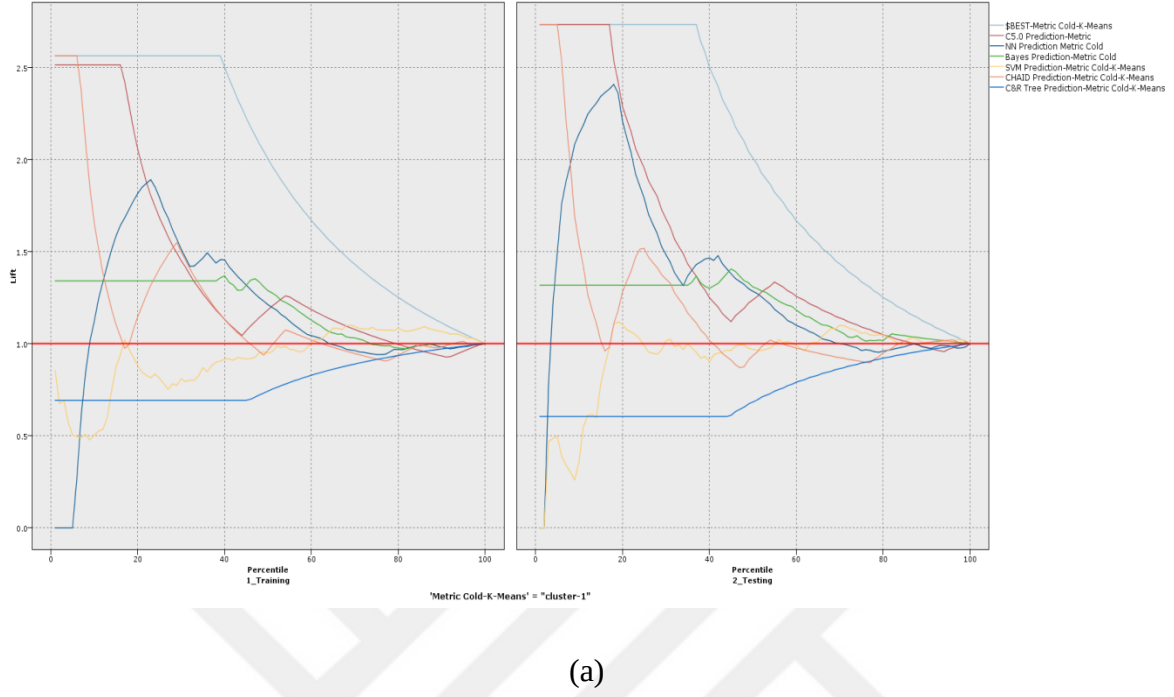


(a)



(b)

EK-12 Şehir içi yakıt tüketimi için (a) yükseltme (b) ROI grafiği (Benchmarking)



ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı : DENİZ, Simge
 Uyuğu : T.C.
 Doğum tarihi ve yeri : 15.01.1991, İskenderun
 Medeni hali : Bekâr
 Telefon : 0 (326) 613 65 66
 e-mail : simge.deniz@gazi.edu.tr



Eğitim

Derece	Eğitim Birimi	Mezuniyet Tarihi
Yüksek lisans	Gazi Üniversitesi /Endüstri Mühendisliği	Devam Ediyor
Lisans	Gazi Üniversitesi /Endüstri Mühendisliği	2014
Lise	İstiklâl Makzume Anadolu Lisesi	2009

İş Deneyimi

Yıl	Yer	Görev
2016	Accentive Heidelberg Services	İş Geliştirme Asistanı

Yabancı Dil

İngilizce, Almanca

Yayınlar

Deniz, S., Gokcen, H., Nakhaeizadeh, G. (2016). Application of data mining methods for analyzing of the fuel consumption and emission levels. *International Journal of Engineering Sciences and Research Technology*, 5(10), 377 – 389.

Hobiler

Türk Sanat Müziği



GAZİ GELECEKTİR..