T.C.
HALİÇ ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
YÖNETİM BİLİŞİM SİSTEMLERİ ANABİLİM DALI /
YÖNETİM BİLİŞİM SİSTEMLERİ TEZSİZ YÜKSEK LİSANS
PROGRAMI

VERİ MADENCİLİĞİNİN İŞLETMELERDE SATIŞ SÜRECİNDEKİ ÖNEMİ: SİGORTA POLİÇE DOLANDIRICILIĞI ÖRNEĞİ

DÖNEM PROJESİ

Hazırlayan
Taha KORKMAZ

Dönem Projesi Danışmanı
Dr. Öğr. Üyesi Altan ALAYBEYOĞLU

İSTANBUL
Haziran 2022

T.C.
HALİÇ ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
YÖNETİM BİLİŞİM SİSTEMLERİ ANABİLİM DALI /
YÖNETİM BİLİŞİM SİSTEMLERİ TEZSİZ YÜKSEK LİSANS
PROGRAMI

VERİ MADENCİLİĞİNİN İŞLETMELERDE SATIŞ SÜRECİNDEKİ ÖNEMİ: SİGORTA POLİÇE DOLANDIRICILIĞI ÖRNEĞİ

DÖNEM PROJESİ

Hazırlayan
Taha KORKMAZ

Dönem Projesi Danışmanı
Dr. Öğr. Üyesi Altan ALAYBEYOĞLU

İSTANBUL
Haziran 2022

01/06/2022

PROJE ETİK BEYANI

Dönem projesi olarak sunduğum "Veri Madenciliğinin İşletmelerde Satış Sürecindeki Önemi: Sigorta Poliçe Dolandırıcılığı Örneği" başlıklı bu çalışmayı baştan sona kadar danışmanım Dr. Öğr. Üyesi Altan Alaybeyoğlu'nun sorumluluğunda tamamladığımı, verileri/örnekleri kendim topladığımı, deneyleri/analizleri ilgili laboratuvarlarda yaptığımı/yaptırdığımı, başka kaynaklardan aldığım bilgileri metinde ve kaynakçada eksiksiz olarak gösterdiğimi, çalışma sürecinde bilimsel araştırma ve etik kurallara uygun olarak davrandığımı ve aksinin ortaya çıkması durumunda her türlü yasal sonucu kabul ettiğimi beyan ederim.

Taha KORKMAZ

ÖNSÖZ

Bilgi teknolojilerinin gelişmesi veri sayısını da arttırmaktadır ancak her veri bilgi demek değildir. Bilgi verinin işlenmiş halidir. Verinin işlenerek bilgiye dönüştürülmesi süreci de veri madenciliğinin doğmasına sebep olmuştur. Bu çalışmada da KNİME ile bir araç sigorta poliçe dolandırıcılığı veri seti üzerinde çalışarak bir sigortalının dolandırıcılık yapıp yapmayacağı konusunda bir kestirim yapmaya çalışılmış, geçmişte dolandırıcılık yapan sigortalıların ortak noktaları belirlenerek en önemli değişkenler çalışmada gösterilmiştir.

Tez çalışması süresince ilgisini hiç esirgemeyen bilgi ve tecrübesini benimle paylaşan değerli hocam ve dönem projesi danışmanım Dr. Öğr. Üyesi Altan Alaybeyoğlu' a saygılarımı sunar ve teşekkür ederim.

Haziran, 2022 Taha KORKMAZ

İÇİNDEKİLER

Sayfa No.

PROJE ETİK BEYANI ..... **i**
ÖNSÖZ ..... **ii**
İÇİNDEKİLER ..... **iii**
KISALTMALAR ..... **v**
ŞEKİL LİSTESİ ..... **vi**
ÖZET ..... **vii**
ABSTRACT ..... **viii**
1. GİRİŞ **1**
2. SİGORTACILIK SEKTÖRÜ HAKKINDA BİLGİ **3**
2.1. Sigortacılık Sektörünün Tarihsel Gelişimi 3
2.2. Sigortacılık ve Riskleri 4
2.3. Literatür Taraması: Sigorta Sektöründe Veri Madenciliği Uygulamaları 5
2.4. Sigorta Sektöründe Veri Madenciliği İhtiyacı ve Bilgi Paylaşımı 8
3. VERİ TABANLARINDA BİLGİ KEŞFİ **9**
3.1. Veri Tabanlarında Bilgi Keşfi Aşamaları 9
4. VERİ MADENCİLİĞİ **11**
4.1. Veri Madenciliği ve Kullanım Alanları 11
4.2. Veri Madenciliğinde Kullanılan Yazılımlar 12
4.2.1. KNIME 12
4.2.2. SPSS Modeller 13
4.2.3. STATISTICA Data Miner 13
4.2.4. ODM (Oracle Data Mining) 13
4.2.5. DBMiner 13
4.2.6. Enterprise Miner 14
4.2.7. Intelligent Miner 14
4.2.8. Weka 14
4.3. Veri Madenciliği Modelleri 15

4.3.1. Tanımlayıcı Modeller 15
4.3.2. Tahmin Edici Modeller 15
5. KARAR AĞAÇLARI 16
5.1. Karar Ağacı 16
5.2. Karar Ağacı Oluşturma 16
5.2.1. Böl ve Elde Et 16
5.2.2. ID 3 Algoritması 18
5.2.3. C4.5 Algoritması 19
5.2.3.1. Bölünme ve Dallanma bilgisi 19
5.2.3.2. Sayısal özellikler 20
5.2.3.3. Kayıp veriler 21
5.3 Ağacın Testi 21
5.4. Ağacın Budanması 22
6. UYGULAMA 24
6.1. Problemin Tanımlanması Aşaması 24
6.2. Veri Tanıma Aşaması 24
6.3. Veri Hazırlama Aşaması 25
6.4. Modelleme Aşaması 27
6.5. Değerlendirme Aşaması 31
6.6. Uygulama Aşaması 31
7. SONUÇLAR 32
KAYNAKLAR 34
ÖZGEÇMİŞ 36

KISALTMALAR

CRM : Müşteri İlişkileri Yönetimi

DMG : Data Minig Group

DMQL : Data Mining - Query Language

KNIME : Konstanz Information Miner

ODM : Oracle Data Mining

OLAP : Çevrimiçi Analitik İşleme

PMML : Predictive Modeling Markup Language

VTKB : Veri Tabanlarında Bilgi Keşfi Aşamaları

Weka : Waikato Environment for Knowledge Analysis

XML : Extensible Markup Language

ŞEKİL LİSTESİ

Sayfa No.

Şekil 3.1. Veri Tabanlarında ilgi Keşfi Aşamaları 10
Şekil 5.1. Örnek Olay Kümesi 17
Şekil 5.2. Örnek Olay Kümesinin Büyüklük Sınıfına Göre Bölünmesi 17
Şekil 5.3. Büyüklük Sınıfına Göre Bölünmesinin Ağacın Bölünmüş Kümeleri Tekrar Biçim Özelliğine Göre Tekrar Bölünmesi Sonucunda Ortaya Çıkan Ağaç 18
Şekil 5.4. Ş Ağacın Bölünmüş Kümeleri Tekrar Biçim Özelliğine Göre Tekrar Bölünmesi Sonucunda Ortaya Çıkan Ağacının Sınıflandırılmamış Olaylarının Renk Özelliği ile Sınıflandırılması 18
Şekil 6.1. Kayıp Veriler Üzerinde Yapılan İşlemler 26
Şekil 6.2. Karar Değişkeni Üzerinde İşlem 26
Şekil 6.3. Renklendirme İşlemi Yapılmadan Önce 26
Şekil 6.4. Renklendirme İşlemi Yapıldıktan Sonra 27
Şekil 6.5. Verinin Bölünmesi 27
Şekil 6.6. Karar Ağacı 28
Şekil 6.7. Karara En Çok Etki Eden Değişkenler 31
Şekil 6.8. Hata Oranı 31

ÖZET

VERİ MADENCİLİĞİNİN İŞLETMELERDE SATIŞ SÜRECİNDEKİ ÖNEMİ: SİGORTA POLİÇE DOLANDIRICILIĞI ÖRNEĞİ

Günümüzde gelişen veri işleme teknikleri ve araçlarının etkisiyle şirketlerin birbiri ile olan rekabetleri farklı bir boyut almıştır. Artan rekabette şirketlerin faaliyetleri sonucu elde ettikleri verileri işlemeleri ve veriden anlamlı bilgi üretmeleri büyük önem arz etmektedir. Bu gerekliliğe veri tabanlarında bilgi keşfi ve veri madenciliği uygulama ve tekniklerinin gelişimi eşlik etmiştir. Veri madenciliğinde amaçlanan, veri tabanlarında saklanan veriler arasındaki gizli kalmış ilişkileri ortaya çıkarmaktır.

Bu çalışma kapsamında da sigortacılık sektörüne dair sigorta sektörünün tarihsel gelişimi, veri madenciliği kullanılarak sigortacılık sektöründe yapılmış uygulamalar ile ilgi literatür taraması, sigortacılık sektöründe kullanılan veri madenciliği uygulamaları ve son olarak da sigorta sektöründe veri madenciliği ihtiyacına, veri madenciliğinin tarihsel gelişimi, veri tabanlarında bilgi keşfinin aşamaları, veri madenciliği hakkında kullanım alanları, veri madenciliği aşamaları, bir veri madenciliği projesi geliştirilirken kullanılabilecek uygulamalar, veri madenciliği modelleri, karar ağaçları, veri madenciliği algoritmaları, dair bilgiler vermektedir. Verilen bu bilgilere ek olarak veri madenciliği kullanılarak yapılan bir uygulama da yer almaktadır. Veri bilimi platformu Kaggle üzerinden erişilen, 1993-1995 yılları arasında Amerika'da toplanan araç sigorta poliçe dolandırıcılığı veri seti üzerinde çalışarak bir sigortalının dolandırıcılık yapıp yapmayacağı konusunda bir kestirim yapmaya çalışılmış, geçmişte dolandırıcılık yapan sigortalıların ortak noktaları belirlenerek en önemli değişkenler çalışmada gösterilmiştir.

Anahtar Kelimeler: *Bilgi, Karar Ağacı***,** *KNIME, Sigortacılık, ,Poliçe Veri Madenciliği,*

ABSTRACT

THE IMPORTANCE OF DATA MINING IN THE SALES PROCESS IN BUSINESSES: EXAMPLE OF INSURANCE POLICY FRAUD

Today, thanks to technology, the competition between companies has increased rapidly. In the increasing competition, it is of great importance for companies to access information, and developing computer technologies have made it possible to store much larger data in databases. Companies also aim to transform raw data into information by processing this data. This necessity has led to the emergence of the concepts of knowledge discovery and data mining in databases. The aim of data mining is to reveal hidden relationships between data stored in databases.

Within the scope of this study, the historical development of data mining, the stages of information discovery in databases, usage areas about data mining, data mining stages, applications that can be used while developing a data mining project, data mining models, decision trees, data mining algorithms, historical information about the insurance industry. development, data mining applications used in the insurance industry, and finally, the need for data mining in the insurance industry. In addition to this information given, there is an application made using data mining. By working on the car insurance policy fraud data set collected in the USA between 1993-1995, an attempt was made to make an estimation of whether an insured would commit fraud, and the most important variables were shown in the study by determining the common points of the insured who committed fraud in the past.

Keywords : *Data Mining, Decision Tree, Information, Insurance, KNIME, Policy*

1. GİRİŞ

Teknolojinin hızla gelişmesi ile veriye ulaşmak, veriyi toplamak, veriyi depolamak kolaylaştı ancak veri tek başına yeterli değildir. Veri bir amaca hizmet edecek şekilde işlenmelidir ki veri arasındaki bağıntılar ve geçmişte gerçekleşen olaylar incelenerek geleceğe dair bir kestirim yapılabilsin. Veri madenciliği de bu görevi üstlenmektedir. Veri madenciliği veri yığınları arasından bilgi elde etmek anlamına gelmektedir.

Veri madenciliği günlük hayatın birçok noktasında büyük önem teşkil ettiği gibi aynı şekilde işletmeler için de büyük önem teşkil etmektedir. Artan rekabet koşullarında işletmeler diğer işletmelere göre avantaj sağlamak zorundadır. İşletmeler veri madenciliği sayesinde avantaj elde edebilirler, veri madenciliği kullanarak pazar analizi, piyasa analizi ve müşteri hareketleri analizi, kampanya, fiyatlandırma, reklam, stok gibi konularda çalışmalar yaparak kendilerine avantaj sağlamayı hedefleyebilirler.

Bu bağlamda çalışmamız giriş bölümü 1. kısım olmak üzere toplamda 7 kısımdan oluşacaktır. 2. kısımda sigortacılık sektörü, sigortacılık sektöründe veri madenciliği ihtiyacı ve sigortacılık sektöründeki riskler hakkında bilgi verilecek. Sektördeki veri madenciliği uygulamalarına dair kısa bir literatür taraması sunulacaktır. 3. kısımda veri tabanlarında bilgi keşfi ve bilgi keşfinin aşamalarına dair bilgiler verilecektir. 4. kısımda veri madenciliğin kullanım alanları ve veri madenciliğinde kullanabileceğimiz programlar, veri madenciliği modelleri ile ilgili bilgiler verilecektir. 5. kısımda karar ağaçları, karar ağacı oluşturma, karar ağacı oluştururken kullanabileceğimiz yöntem, algoritmalar ve oluşturduğumuz ağacın testi ve budanması ile ilgili bilgi verilecektir. 6. Kısımda ise verilen tüm bu bilgiler kullanılarak veri bilimi platformu Kaggle üzerinden erişilen 1993-1995 yılları arasında Amerika'da toplanan araç sigorta poliçe dolandırıcılığı veri seti üzerinde KNIME isimli program ile çalışarak veri madenciliği uygulaması hayata geçirilecektir. Bu kısımda ilk aşama problemi belirlemektir. Ardından veri seti içerisindeki değişkenler hakkında bilgi verilecek, değişkenler üzerinde filtremeler yapılacak, kirli ve gürültülü

veriler azaltılacak, modelimizin hata oranını belirleyebilmek için veri setimiz %70 eğitim, %30 test kümesi olmak üzere ikiye bölünecektir. Bir sonra ki aşamada ise veri madenciliği teknikleri uygulanarak karar ağacı oluşturulacak ve karar vermemize en çok yardımcı olacak değişken belirlenecektir. Diğer değişkenler de önem sırasına göre listelenerek bir sigortalının dolandırıcılık yapıp yapmayacağı konusunda bir kestirim yapılmaya çalışılacaktır. Oluşturulan modelin de doğruluğu eğitim kümesi üzerinden öğrenilerek test kümesi üzerinden test edilecek ve sonuç paylaşılacaktır. Son olarak 7. kısımda ise genel olarak çalışmadan çıkarılabilecek sonuçlar paylaşılacaktır (URL-1).

2. SİGORTACILIK SEKTÖRÜ HAKKINDA BİLGİ

Sigorta sözcüğünün kökü "sicurta" olarak İtalyancadan gelmektedir. Türkçede ise daha önce "sigcuriye", "sikorta", "sikurita" "sikurta" ve "sigurta" gibi sözcükler kullanılmış ve son olarak "sigorta" kelimesi ortaya çıkmıştır. Sigortanın kelime anlamı olarak, "güvence" olarak bilinmektedir. Kavram olarak ise sigorta, bir insanın hayatı boyunca karşılaşabileceği muhtemel kötü olayların yaratabileceği hasarların sigortalayan ve sigortalı arasındaki anlaşmaya bağlı kalarak giderilmesine ile ilgili bir faaliyettir. İnsanlar, belirsizliklerin, risklerin olduğu bir dünyada yaşamaktadır. İnsanlar kendilerini güven altında hissetmek için bu riskleri azaltmak isterler sigorta, bu riskleri azaltma ve gerçekleşen risklerin verdiği maddi zararı en aza indirme veya mümkünse tamamını gidermek amacıyla doğmuştur (Yayla, 2019, s. 108-109).

2.1. Sigortacılık Sektörünün Tarihsel Gelişimi

Tarihte sigortacılık olarak adlandırabileceğimiz ilk uygulamalara yaklaşık 4000 yıl önce Babiller' de rastlanılmıştır. Zamanın ticaret merkez olarak bilinen Babil' de, kervan tüccarlarına borç veren zenginler, kervanların yolda bir sorun ile karşılaşmasına karşın tüccarların adlıkları borçları silmekte ve buna karşılık borcu tüccarlardan geri aldıkları zaman, taşıdıkları riskin karşılığı olarak ana borç miktarı üzerinden belirli bir oranda para almaktaydılar. Bu olay daha sonralarda ise Kral Hammurabi tarafından yasallaştırılmıştır. Hammurabi Kanunlarının en büyük özelliği haydutların saldırısına uğrayan kervanların zararlarının bütün diğer kervanlar arasında paylaşılmasını öngörmeseydi. Bu, tehlike paylaşmasının kara taşımacılığındaki ilk sigorta örneği olarak gösterilir.

Prim bazlı sigorta yaklaşık M.S. 1250'lü yıllarda Venedik, Floransa ve Cenova kentlerinde görülmüştür, sigortanın günümüzdeki halini alması ise 14. yy'in başlarıdır. Ekonomik şartların değişiklik göstermesi ile ticaret, 14. yy'in başlarından itibaren hızlı bir şekilde gelişim gösterdi. O zamlanalar deniz üzerinde taşımacılığın en sık yapıldığı ülkelerden İtalya' da deniz sigortası kavramı ortaya çıkmıştır. Tarihte ilk sigorta poliçesi olarak kabul gören sözleşme 23 Ekim 1347 tarihinde İtalya'nın Cenova

kentinde bulunan limandan Mayorka'ya, "Santa Clara" adlı geminin yükünü temin etmek amacıyla imzalanmıştır. İlk sigorta şirketi ise 1424 yılında, yine İtalya'nın Cenova kentinde kurulmuştur. Sigortacılık konusunda ilk resmi gelişme de 1435 yılında yayınlanan Barselona Fermanı olarak bilinmektedir. Ardından İtalya'daki başlangıçtan sonra, deniz sigortaları özellikle 18. yy' da İngiltere'de de yaygın hale gelmiştir.

İlk olarak denizde başlayan sigortacılık, ilerleyen zamanlarda hayat sigortası kavramının geliştirilmesinin önünü açtı. Gemi ve yükün sigorta edilebilmesi, yolcular, tayfalar ve kaptanın da sigorta edilebileceği düşüncesini akıllara getirdi. 17.yy.'da İtalya'da yaşayan bir banker olan Tonti'ni tarafından geliştirilen "Tontines" adı verilen sistemde, belirli kişiler toplanarak, belirlenen bir süre için ortaya belirli bir para koymakta ve sürenin sonunda yaşamına devam eden kişiler parayı aralarında pay etmekteydi. İnsanlar genellikle, kendilerinin diğer kişilerden daha fazla yaşayacaklarını düşündükleri için ilgi gören bu sistemde ölenlerin maddi açıdan zarar gördükleri düşünülerek, öngörülen süreden önce ölenler için de ölüm rizikosu karşılığı prim ödenmesi öngörüldü ve hayat sigortalarına başlangıcı da bu şekilde olmuştur.

Sigorta tarihinde bir başka önemli olay ise 1666 yılında Londra'da, dört gün boyunca süren, 100 kilisenin ve 13000 ev yanmasına neden olan yangıdır. Yangın yalnızca İngiltere'de değil tüm dünyada yankı buldu ve insanların canlarının, mallarının korunması fikri ile yangından 1 yıl sonra da İngiltere'de "yangın bürosu" kurulmuştur.

20. yüzyılın başlarında ise artık sigorta şirketleri birçok türde sigorta ihtiyacına yanıt verebilecek seviyeye erişmiş kuruluşlar olarak etkin şekilde hizmet verebilecek düzeye bir gelmiştir ve sigortacılık sektörü, kavramı günümüzdeki halini almaya başlamıştır (Altun, 2007, s. 4-7).

İlerleyen yıllarda ise veri madenciliği sigortacılık sektörünün içerisine girmiştir. Veri madenciliği sayesinde sigorta sistemleri çeşitli karar destek sistemleri oluşturmuş, çeşitli fiyatlandırma politikaları ve reklam kampanyaları gerçekleştirmiştir.

2.2. Sigortacılık ve Riskleri

Risk kavramı sigortacılıkta, kaybetme tehlikesinin varlığı, belirsizlik ve beklenenin dışında bir sonuçtan başka herhangi bir sonucun ortaya çıkma olasılığı

olarak ifade edilir. Sigortacılıkta risk işletme riskleri ve sistematik riskler olmak üzere iki ana başlık altında incelenir ve bu ana başlıklarda alt başlıklara bölünür.

Sigorta şirket tarafından yapılan yanlış satış, dolandırıcılık, genel olarak ihtiyaç duyulan tutarda fonun gerektiği zamanda makul bir maliyetle bulunamaması, elde var olan finansal varlığın istenilen zamanda ve fiyatta elden çıkarılamaması veya transfer edilememesi, özellikle hayat sigortalarında sabit maliyetlerinin prim gelirleri ile finanse edilmesi sözleşmelerin feshedildiği taktirde sabit harcamaların aksamasına sebep olur, masraflarını karşılayamayacağı tutarlarda fazla riskli varlıklar için poliçe düzenlenmesi, sigorta şirketleri bazı durumlarda sermayelerini korumak için döviz ve değerli madenlere yatırım yaparlar bu da piyasanın durumuna göre sermaye kayıplarına sebep olabilir, sigorta şirketleri yine sermayelerini korumak için sermayelerini faize yatırabilirler faize yatırılan şirkette yaşanabilecek sorunlar, sigorta şirketleri kendi varlıkları yanı sıra piyasadan topladıkları fonları da değerlendirdiği için piyasada olacak düşüşler, küresel veya yerel çapta olabilecek bir ekonomik kriz sigortacılık sektöründe görülen risklerdendir (Sarıoğuz, 2007, s. 35-49).

2.3. Literatür Taraması: Sigorta Sektöründe Veri Madenciliği Uygulamaları

Bu başlık altında geçmişte yapılan sigortacılık sektörü ve sigortacılık sektöründe veri madenciliği kullanım alanları, sigortacılık sektöründe veri madenciliği kullanım gereksinimi ile ilgili daha önce yapılmış araştırmalar ve sigortacılık sektöründe kullanılan veri madenciliği uygulamaları hakkında bilgi hakkında bilgi verilecektir.

2021 yılında sigortacılık sektöründe makine öğrenmesi ile müşteri kaybının nasıl önüne nasıl geçilebileceği ile ilgili bir çalışma yapılmıştır. Geçmişte sigorta şirketini bırakan müşterilerin bilgileri incelenerek mevcutta bulunan müşterilerin sigorta şirketini bırakma ihtimalleri belirlenmiş ve bir karar ağacı oluşturulmuştur (Akyiğit, 2021).

2018 yılında sigorta sektöründe veri madenciliğinin kullanım alanları ile ilgili yazılan dergi makalesinde veri madenciliğinin sigorta şirketleri için piyasayı analiz etme, fiyatlarını belirleme ve sigorta suiistimalleri gibi konularda yardımcı, olduğu çalışmanın amacının ise sigorta sektörü gibi verileri doğru değerlendirmenin ve yorumlanması gereken bir sektörde veri madenciliği uygulamalarının gereksinimi vurgulanmıştır (Akpınar, 2018).

2009 yılında sigortacılık sektöründe risk analizi ile ilgili çalışılma yapılmış ve bir veri madenciliği uygulaması ile desteklenmiştir. Çalışmada sigortacılık sektörünün adımlarından biri olan hasar ihbarlarının olumsuz sonuçlanması ile ilgili bağıntılar belirlenerek ve yeni ihbarların sonucuna dair kestirim yapılmaya çalışılmıştır. Olumsuz sonuçlanmasına sebep olabilecek risklerin belirlenmesinde veri madenciliği yöntemlerinden biri olan karar ağaçlarından faydalanılarak veri madenciliği kullanarak sigorta sektöründe bir çalışma yapılmıştır. (Muslu, 2009).

2013 yılında yapılan bir çalışmanın veri madenciliği yöntemlerinin sigortacılık sektöründe müşteri ilişkilerinin yönetimi ile ilgili kullanılmasına ilişkin uygulama gerçekleştirilmesi hedeflenmiştir. Bu amaçla bir sigorta şirketlerinin müşterilerine ait bir veri seti kullanmış olup ve buradaki verilerinden yararlanılarak birliktelik kuralı analizi ve kümele analizi gerçekleştirilmiştir. Bu analizlerin amacı, müşterilerin ürün satın alma alışkanlıklarının bulunması ve benzer müşteri kümelerinin ortak özelliklerinin elde edilmesidir. Bu sonuçlar yardımıyla, şirketlerin müşterilerini tanıması ve yeni pazarlama stratejileri geliştirmeleri hedeflenmiştir. (Bahar, 2013)

2007 yılında sigortacılık sektöründe müşteri ilişkileri yönetimine dair veri madenciliği kullanılarak bir araştırma yapılmıştır. Çalışmanın ilk bölümünde sigortacılık ve müşteri ilişkileri hakkında bilgi verilmiş olup yapılan teknik analizlerin amaçları vurgulanmıştır, müşteri sadakati, müşteri memnuniyeti vb. tanımları detaylı bir şekilde açıklanmıştır. İkinci bölümde veri madenciliğine dair bilgiler vermiş, burada kullanılan yöntemler hakkında bilgi vermiştir. Üçüncü bölümde ise bir veri setinde veri madenciliği teknikleri uygulanarak işletmelerin müşterileri hakkında daha çok bilgi edinerek müşterilerine özel satış kampanyaları, özel indirimler belirlenmiştir. Müşteriler belirli segmentlere bölünmüştür. (Kasap, 2007)

2021 yılında yapılan bir başka çalışma ise sigortacılık sektöründe siber güvenlik yönetimi ve riskin azaltılmasında siber güvenlik sigortalarının rolü ile ilgilidir. Çalışmanın ilk bölümünde sigortacılık hakkında temel bilgiler verilmiştir. İkinci bölümde günümüzde ve gelecekte firmalar ve hatta ülkeler için bile büyük tehdit oluşturacak siber tehdit, siber risk, siber saldırı hakkında bilgi verilmiş olup örnekler ile pekiştirilmiştir. Üçüncü bölümde ise elde edilen bulgular dahilinde, Türkiye'de ve dünyada siber risk sigorta faaliyetlerinin riskin azaltılmasındaki etkin rolüne ve önerilere sonuç bölümünde yer verilmiştir (Karadağ, 2021).

2021 yılında yapılan bir diğer çalışma ise sigortacılık sektöründe dijitalleşme ile ilgilidir ve çalışma içinde sigorta sektöründe dijitalleşme ile ilgili müşteri ve acente

çalışanları üzerine bir uygulama yapılmıştır. Çalışmanın birincil amacı sigorta müşterilerinin ve sigorta sektöründeki acente çalışanlarının dijitalleşmeye yönelik tutumlarını ölçmek olarak belirlenmiştir. Ankara ilinde faaliyet gösteren bir sigorta acentesi çalışanları ile sigorta müşteriler üzerinde bir anket çalışması yapılmıştır. Araştırmanın sonucunda, dijital dönüşümün bir seçenek değil mecburiyet olduğu görüşü paylaşılmıştır (Varol, 2021).

Sigorta sektöründe, veri madenciliği, sektörün gelişmesi ve bilgi elde etme bakımından büyük önem teşkil etmektedir. Sigortacılık sektörü ile ilgili teknik işlemlerin tamamı, istatistiksel bilgiler, matematiksel modeller ve yöntemlerin kullanılması ile gerçekleştirilmektedir. Örnek verecek olursak; hayat sigortası üzerine faaliyet gösteren şirketler, ölüm tablolarını hazırlarken, geçmiş yıllardaki verileri baz alarak kişilerin ortalama yaşam süreleri ile ilgili çalışmalar yapar ve yeni bir ürün çıkaracakları zaman bu çalışmalar neticesinde ulaşılan bilgileri kullanmaktadır. Hayat dışı sigorta şirketleri için de aynı durum geçerlidir. Bu çalışmalar sonucunda ortaya çıkan risk profillerine göre fiyatlandırma yapılmaktadır. Sigortacılık sektöründe veri madenciliği uygulamaları şirketlere; mevcutta müşteri olmayan yeni poliçe talep edebilecek potansiyel müşterilerin tahmininde, sigorta suiistimallerinin belirlenmesi ve önlenmesinde, risk doğurabilecek müşterilerin tespit edilmesi noktalarında fayda sağlamaktadır.

Yapılan çalışmalarda sigorta poliçesi satın alan müşterilerin yıllık olarak önemli bir kesimin poliçelerini tekrar yenilemediklerini görülmektedir, başka bir ifade ile mevcut müşteriyi ellerinde tutamıyorlardır. Bu durumda sigorta şirketleri yeni sigorta ürünleri geliştirmesi ve iş hacmini büyütmeyi düşünmesi gerekirken mevcut müşterilerini de kaybetme ihtimaline bir çözüm bulunması gereken önemli bir sorun olarak görülmektedir. Genellikle müşteriler psikolojik olarak, ödedikleri tutar karşılığında kaliteli ürün veya hizmet almak istemektedirler. Sigortacılığın en zor yanlarında biri bu noktada önümüze gelmektedir. Sigorta şirketleri tarafından güvence altına alınan riskler gerçekleşmediği ve hasar olmadığı müddetçe müşteriler sigortacılık hizmetinden doğrudan faydalanamamaktadır ve bu nedenle müşteriler poliçelerini yenileyip yenilememe ikilemine düşerler. Sigorta şirketlerinin de mevcutta bulunan müşterilerini sistemleri içerisinde tutmak, belirli zamanlarda değerlendirmek, talep ve ihtiyaçlarına göre bazı özel avantajlar geliştirerek sunmaları gerekmektedir. Mevcut sigorta poliçelerini inceleyip, gruplandırarak, bu gruplara özel yeni ürünler geliştirmek, müşterilerin davranışlarını ek hizmetler sunabilmek için takip etmek,

kampanya yönetimi yapmak, müşterilerin alışkanlıklarına duyarlı olmak veri madenciliği uygulamaları ile yapılabilmektedir (Akpınar, 2018, s. 112-113).

Literatür çalışmasının ve sigortacılık sektöründe veri madenciliğinin kullanım alanlarının incelenmesinin ardından veri madenciliğinin sigortacılık sektöründe, müşteri kazanımı, müşteri kaybını önleme, müşteri memnuniyetini arttırma, müşteri suiistimallerini en az orana indirme gibi alanlarda kullanılabileceği görülmüştür. Veri madenciliğinin yanı sıra sigortacılık sektöründe teknoloji kullanımı gereksinimi olduğu görülmüştür.

2.4. Sigorta Sektöründe Veri Madenciliği İhtiyacı ve Bilgi Paylaşımı

Veri madenciliği uygulamalarının verimli ve etkin şekilde kullanılması şirketlerin bilgiyi yönetmesine imkan vermektedir. Sigorta türlerine göre değişkenlik göstermekle birlikte, yüksek seviyelerde haksız gelir sağlamak amacı ile sigorta suiistimalleri görülmektedir. Veri madenciliği uygulamaları ile elde edilecek bilgiler ile suiistimallerin önüne geçilerek edilecek tasarruflar, ciddi tutarlara ulaşabilir.

Sigorta şirketleri poliçe primlerini belirlemek veya piyasa ve pazar yapısını değerlendirmek için piyasadaki mevcut veri üzerinde veri madenciliği uygulamalarını kullanabilmektedirler. Verilerini verimli bir şekilde kullanamayan sigorta şirketlerinin, operasyonel olarak kullandıkları sistemlerin de yeni gelişen teknolojilere adapte olamadığı görülmektedir. İyi bir veri yönetimi uygulamasının sigorta şirketlerine, sektörde önemli avantajlar sağlayacağı düşüncesi ile hareket eden şirketler, bu öngörü ile diğer sigorta şirketlerine göre rekabet konusunda önemli bir avantaj sağladıkları görülmektedir.

Veri madenciliği sigorta şirketlerine mevcut müşteri hakkında daha çok bilgi vermektedir, mevcut müşterileri gruplandırarak ortaya çıkacak risk davranış modellerinin yeni başvuruda bulunan müşterilere uygulanmasını sağlayarak, riski minimuma indirme, mevcutta bulunan müşterilerinin ödeme performanslarını inceleyerek, düşük ödeme performansına sahip müşterilerin ortak özelliklerinin belirlenmesi ve benzer özelliklere sahip tüm müşteriler için yeni risk yönetim politikaları oluşturulması veri madenciliği sayesinde yapılabilmektedir (Akpınar, 2018, s. 113-115).

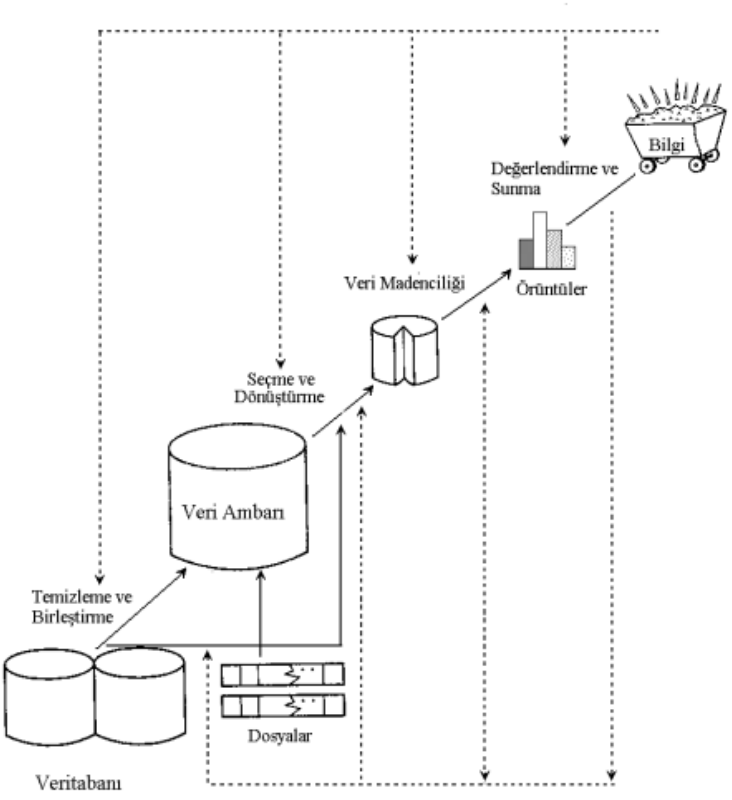
3. VERİ TABANLARINDA BİLGİ KEŞFİ

Geçmişte veri madenciliği ve bilgi bulma daha kolayken günümüzde birçok sistemdeki veri miktarı terabayt boyutundan daha büyük hale geldi ve günümüzde kullanılan modern veri madenciliği algoritmalarının, tekniklerin geliştirilmesinde rol oynadı. Veri tek başına bir şey ifade etmezken veri, veri madenciliği sayesinde bilgi haline dönüştürüldü ve insanlar verinin içinden anlamlı, değerli, önceden bilinmeyen bilgiyi keşfetmeye başladı. (Koçtürk, 2010, s. 5)

Veri tabanlarında bilgi keşfi çok disiplinli birbirinden bağımsız faaliyetleri içerir. Bu, veri depolama ve veriye erişimi, algoritmaları, büyük veri kümelerinde ölçeklendirmeyi ve sonuçları yorumlamayı kapsar. Veri seti içerisinde, veri temizleme işleminin yapılması ve veriye erişim sürecinin hızlandırılması veri tabanlarında bilgi keşfi sürecini kolaylaştırır. Veri seti içerisinde açığa çıkarılan örüntüler yeni veriler üzerinde de geçerli olmalı ve bir dereceye kadar kesinliğe sahip olmalıdır. Açığa çıkarılan bu örüntüler yeni bilgi olarak kabul edilmektedir (Gürel, 2019, s. 6).

3.1. Veri Tabanlarında Bilgi Keşfi Aşamaları

Veriden bilgi elde etme süreci küçük adımlara bölünmüştür. İlk adım verimizi seçmektir. Elimizde aynı amaca hizmet eden birden fazla veri seti olduğunda üzerinde sorgulama yapabileceğimiz bir veri seti seçmeliyiz hatta gerekirse iki farklı veri setini birleştirerek tek bir veri seti oluşturabiliriz. Ardından veri işleme adımı gelir, bu adımda seçilen veri setini bulunan hatalı verilerden arındırırız ve veriyi dönüştürürüz, eksik verileri tamamlarız. Örnek olarak, integer olarak tanımlanmış bir ve sıfır değerlerinden oluşan bir değişkeni boolean veri tipine çevirerek doğru, yanlış, evet, hayır şeklinde veriyi daha okunur bir hale getirebilir. Daha Sonra veri indirgeme adımı gelir, veri indirgeme aşmasında ilgisiz, niteliksiz, tekrarlayan verilerin çıkarılması adımıdır. Bu işlemde verinin boyutu azaltılarak veri madenciliği teknikleri uygulanırken çalıştırılacak sorguların daha hızlı sonuç üretmesi amaçlanır. Ardından veri madenciliği süreci başlar, bu adımda veri madenciliği yöntemleri ve algoritmaları uygulanır ve son adıma geçilir son adımda ise elde edilen modelin yenilik, yararlılık ve basitlik kıstaslarına göre değerlendirilmesi yapılır (Şen, 2008, s. 4-7).



Şekil 3.1. Veri Tabanlarında ilgi Keşfi Aşamaları

4. VERİ MADENCİLİĞİ

4.1. Veri Madenciliği ve Kullanım Alanları

Veri madenciliği, günümüzün en güncel ve en çok kullanılan teknolojilerden birisidir. Bilgisayar sistemlerinin her gün hızla gelişiyor ve güçlerinin artıyor olması, veri tabanlarında çok daha büyük boyutlarda verilerin depolanabilmesi imkanı tanımaktadır. Veri madenciliği de saklanan veri içerisinden, veri analizi ve ileri seviyede matematiksel algoritmalar kullanılarak, veri içerisindeki örüntüleri ve yönelimleri keşfederek ileride yaşanabilecek olayların sonuçlarını değerlendirmek için büyük veri kümeleri arasındaki sıralama işlemidir (Uluyardımcı & Zontul, 2020, s. 4).

Veri madenciliği ile ilgili yapılan iki tanım aşağıda verilmiştir.

Veri madenciliği büyük boyutlarda veri içerisinde gelecekte yaşanması muhtemel olaylar ile ilgili kestirim yapmamızı sağlayabilecek örüntü ve kuralların bilgisayar programları sayesinde aranmasıdır. (Alpaydın, 2000, s. 1)

Veri madenciliği, veri içerisinde yapılan işlemler sonrasında veriler arasında ki ilişkilerin açığa çıkarılması amacıyla bir algoritma çalıştırma işlemidir (Zaimoğlu, 2018, s. 13).

Veri madenciliği finans, bankacılık, sigortacılık, yazılım gibi birçok sektör de kullanılmak ile birlikte gelecekte de hayatımızın önemli bir parçası olacak otonom araçların olası bir kaza durumunda kazayı önlemek için vereceği tepki, metinin bir kısmında kişinin olumlu veya olumsuz yönde duyguya sahip olduğunun belirlenmesi ve sepet analizi olarak da bilinen müşterilerin alacakları potansiyel ürünlerin, ürünlerin raf sırasının belirlenmesinde, online alışveriş sırasında ürünlerin önerilmesinin de arkasında veri madenciliği vardır.

Bir veri madenciliği uygulamasının geliştirilmesi genel olarak 6 aşamaya bölünmüştür.

1.Problemin Tanımlanması Aşaması

Veri madenciliğinin en temel aşamasıdır. Problem araştırılır, tanımlanır ve üzerinde çalışmaya başlanır.

2.Veri Tanıma Aşaması

Veriyi bir araya getirmek ile başlar verilerin tipleri ve nitelikli veriler tanımlanır, veri içerisindeki değişkenlerin neyi ifade ettiği araştırılır.

3.Veri Hazırlama Aşaması

Eksik, kirli, gürültülü veriler temizlenebilir veya çeşitli yöntemler ile boş değer yerine uygulanan yöntem sonucunda belirlenen değer yazılabilir.

4.Modelleme Aşaması

Veri tipimiz ve amaçlarımız doğrultusunda uygun modelin oluşturulma aşamasıdır.

5. Değerlendirme Aşaması

Yapılan model üzerinden artık sonuçlar almaya başlanır. Alınan sonuçlar tatmin edici olmadığı taktirde modelleme hatta veri hazırlama aşamasına geri dönülerek tekrar veriler üzerinde çalışarak daha iyi değerler almak hedeflenir.

6.Uygulama Aşaması

Son aşama olarak değerlendirme aşamasında tatmin edici değerler alındı takdirde uygulama hayata geçirilerek kullanılmaya başlar

4.2. Veri Madenciliğinde Kullanılan Yazılımlar

Veri madenciliği için kullanılan birçok yazılım vardır. Yazılımlar işlemlerimizi kolaylaştırmayı amaçlar ve bize sunulan görsel arayüz sayesinde veri özetleme, değişkenlerin analizi, sapmaların tespiti, kümeleme, karar ağacı oluşturma gibi işlemler yapabiliriz. Programların bazıları ücretiz ve açık kaynak kodlu iken bazıları lisans ücreti karşılığında kullanıma sunulmuştur.

4.2.1. KNIME

İsmi Konstanz Information Miner'ın kısaltmasından gelmektedir ve aynı ismi sahip şirket tarafından açık kaynak kodlu olarak geliştirilmektedir. Kodlamaya gerek kalmadan görsel olarak bir uygulama geliştirilebilir. KNIME Hub üzerinde diğer KNIME kullanıcıları ile iletişime geçilerek sorular sorulabilir, çözüme ulaşılabilir. Java programlama dili kullanarak ve Eclipse tabanı üzerine kurulmuştur. İçerisinde dahili olarak gelen KNIME Extensions ile uygulamaya yeni eklentiler eklenerek kullanım alanı genişletilebilir. KNIME resmi web sitesi üzerinden ücretsiz bir şekilde indirilerek kullanılabilir. Çoğunlukla CRM, iş zekâsı uygulamaları ile kullanımı yaygındır.

4.2.2. SPSS Modeller

Veri bilimciler için yapılan işleri hızlandırmak için IBM firmasının bünyesinde geliştirilmiş bir yazılımdır. Uçtan uca bir veri manipülasyonu ve veri madenciliği akışları kurabilmemizi sağlamaktadır. Çalışma mantığı genel olarak sürükle ve bırak şeklindedir içerisinde dahili olarak gelen veri madenciliği algoritma çeşitliliği fazladır. Lisans ücreti ödenerek kullanılması gerekir, veri madenciliği algoritması bakımınındın geniştir, kullanım zorluğu ise orta derecededir (Koçtürk, 2010, s. 35).

4.2.3. STATISTICA Data Miner

StatSoft firmasının bir ürünü olan uygulama veri madenciliği sürecinde olan tüm işlemleri kolaylaştırır. Diğer uygulamalara göre daha fazla görsel ara yüze sahiptir, model çıktısı için çeşitli seçenekleri vardır, Visual Basic programlama dili kullanılarak geliştirilmiştir (Koçtürk, 2010, s. 35-36).

4.2.4. ODM (Oracle Data Mining)

Eski adı Darwin olan uygulama Oracle firmasının bir ürünüdür, veri hazırlama, model değerlendirme ve model puanlama süreci boyunca rehberlik etmek için çeşitli sihirbazların kullanılmasını sağlar. Veri analistleri, verileri dönüştürürken, modeller oluştururken ve sonuçları yorumlarken Oracle Data Miner, veri madenciliği adımlarını entegre bir veri madenciliği uygulamasına dönüştürmek için gereken kodu otomatik olarak oluşturabilir. Tüm Oracle Veri Madenciliği işlevlerine, PL/SQL ve/veya Java API'leri tarafından erişilebilir, böylece Oracle Veri tabanınızın üzerinde kurumsal iş zekası uygulamaları geliştirebilirsiniz. ODMimer, Oracle Data Miner modellerinden ve sonuçlarından Java kodu üretir, ODM, modellere tek kullanıcılı birden fazla oturumlu erişim sağlar. ODM programları Java ara yüzünde asenkron veya senkronize olarak çalışabilir. PL/SQL arabirimini kullanan ODM programları eşzamanlı olarak çalışır; PL/SQL'i eşzamanız çalıştırmak için Oracle Zamanlayıcı'nın kullanılması gerekmektedir. (URL-2)

4.2.5. DBMiner

Kanada'da bulunan Simon Fraser Üniversitesi tarafından geliştirilen program çevrimiçi analitik işleme (OLAP) yeteneğini ve Veri madenciliği algoritmalarını birlikte kullanabilmesi en önemli özelliğidir. Diğer uygulamalara göre daha basit bir arayüze sahiptir bu nedenle kullanım zorluğu kolay seviye olarak tanımlanabilir.

Uygulama genel amaçlı geliştirilen DMQL sorgulama dilini kullanır. Bu dili SQL'e benzetebiliriz. DMQL kullanarak çevrimiçi sorgular OLAM ya da OLAP modülüne yönlendirilerek gerçekleştirilebilir (Haberal, 2007, s. 43).

4.2.6. Enterprise Miner

SAS firması tarafından geliştirilen bir veri madenciliği uygulamasıdır. Enterprise Miner yapay sinir ağları, karar ağaçları, 2-aşama modelleri (two-stage models), regresyon analizi, kümeleme, ilişkilendirme,zaman serileri, vb. grafiksel ara yüzü ile kullanım kolaylığı sağlar ve kullanıcılar uygulamanın karmaşıklığı ile uğraşmadan şekilde yalnızca girdi ve çıktılara yoğunlaşabilirler. İstemci bilgisayardaki yazılım gereksinimi Windows 98, 2000 ve NT'dir. Sunucu bilgisayardaki yazılım gereksinimi ise Windows 98, 2000 ve NT ile Linux'dür (Haberal, 2007, s. 47).

4.2.7. Intelligent Miner

Intelligent Miner, IBM tarafından geliştirilen veri madenciliği uygulamalarında kullanılabilecek bir üründür. Birçok algoritma ve görüntüleme aracını içerisinde barındırır. Ayrıca Data Minig Group (DMG) tarafınca tanımlanmış olan Predictive Modeling Markup Language (PMML) için de veri madenciliği modelleri üretmiştir. PMML dosyaları, ilişki modellerini ve düzenlenmiş veri setlerinin istatistiklerini içeren bir XML dosyasıdır (Kavurakçı, Gürkaş Aydın, & Şamlı, 2011, s. 5).

4.2.8. Weka

Uzun adı "Waikato Environment for Knowledge Analysis", Weka, makine öğrenimi amacıyla Waikato Üniversitesi tarafından geliştirilmiş yazılımın ismidir. Kullanım alanı genellikle makine öğrenimi metotlarını ve algoritmalarını içermektedir. Java dilini kullandığı için ve kütüphanelerinin .jar formatı halinde geliyor olması sayesinde, Java programlama dili ile yazılan projelere kolayca uyum sağlaması kullanıcıları için önemli bir özelliktir. Tamamıyla modüler bir yapıda olduğu ve içerisinde barındırdığı özelliklerle veri analizi, veri kümeleri üzerinde görselleştirme, iş zekâsı uygulamaları, veri madenciliği gibi işlemlerini gerçekleştirebilmektedir (Gündüz, 2020).

4.3. Veri Madenciliği Modelleri

Veri madenciliği, verinin üzerinde analiz yapılarak bilgi haline getirilmesi ile ilgilenir. Analizin yapılması için de kullanılabilecek iki yöntem vardır. Bunlardan birincisi doğrulamaya dayalı yöntem diğeri ise keşfetmeye dayalı yöntemdir. Yöntemler veriler üzerinde nasıl analiz yapılarak bilgiye nasıl ulaşılacağını bildirir.

Keşfetmeye dayalı yöntemde veri tabanı içerisinde bulunan mevcut veriler üzerinde işlem yaparak bilginin elde edilmesine yardımcı olur. Yani yeni bilgilerin tanımlanması (Descriptive) veya tahmin (Predictive) edilmesini sağlar. Doğrulamaya dayalı yöntem ise günümüzde pek fazla kullanılmamakla birlikte yeni bir bilgi üretmez mevcut veri tabanları üzerinde basit istatistikler, raporlar ve sorgular oluşturmak için kullanılır (Gürel, 2019, s. 18-19).

4.3.1. Tanımlayıcı Modeller

Karar verme konusunda mevcut verilerdeki örüntülerin tanımlanmasını sağlayan modellerdir. Genel olarak sepet analizi, belirli bir gelir seviyesinde bulunan ve çocuğa sahip aileler ile daha düşük gelir seviyesinde, çocuk sahibi olmayan ailelerin alışveriş alışkanlıklarının farklı olduğunun belirlenmesi tanımlayıcı model için bir örnektir. Tanımlayıcı modeli referans alan modeller, Ardışık, Kümeleme (Clustering), Zamanlı Örüntüler (Sequential Pattern) ve Birliktelik Kuralları (Association Rules) Modelleridir (Koçtürk, 2010, s. 12).

4.3.2. Tahmin Edici Modeller

Sonuçlanmış bir olayı referans alarak bir model geliştirilmesinin ve bu modeli sonucu bilinmeyen veriler üzerinde uygulayarak bir kestirim elde edilmesinin hedeflendiği modeldir. Örnek olarak, bir bankada geçmiş dönemlerde kredi kullanan müşterileri verileri ellerinde bulunur. Bu verilerde kişisel bilgiler bağımsız değişkenler olurken kredinin geri ödenip ödenmediği bağımlı değişkendir.

Bu veriler üzerinden oluşturulan model yardımı ile ileri dönemlerde kredi talebinde bulunan müşterilerin krediyi geri ödeyip ödemeyeceğinin kestirimi yapılabilir. Tahmin edici model tabanlı geliştirilen modeller ise Sınıflandırma (Classification) ve Regresyon (Regression) Modelleridir. (Gürel, 2019, s. 18).

5. KARAR AĞAÇLARI

5.1. Karar Ağacı

Karar ağacı, veriyi sınıflandırma ve öngörüde bulunmak için oldukça sık kullanılan bir veri madenciliği tekniğidir. Karar ağaçları, ağaç şeklinde sınıflandırıcı bir yapıdır ve ağaçtaki her bir düğüm ya bir yaprak düğümü ya da karar düğümünü simgeler. Yaprak düğümü hedef niteliğinin değeridir. Karar düğümü de bir nitelik için uygulanacak olan testin değeridir, bu düğümü, o niteliğe muhtemel olan tüm olası nitelik değerleri takip eder, bu değerler ise karar ağacın dallarını oluşturur. Başka bir ifade ile, düğümler soruları ifade eder, dal bu soruların cevaplarını ve yaprak ise kararın verildiği sınıfı ifade eder. Ağacın birinci düğümü ile sorular sorulmaya başlar ve dalları olmayan düğümler ya da yapraklara gelene kadar bu sorular sorulmaya devam eder. Karar ağacının if – then yapıları gibi kurallarla ifade edilmesi basit bir işlemdir. Bu aşamaya gelene kadar kaç tane yaprak varsa o kadar kural oluşmuştur (Koçtürk, 2010, s. 25).

5.2. Karar Ağacı Oluşturma

Karar ağaçlarının oluşturulması ile ilgili çeşitli yöntemler vardır ancak ağacı oluşturma ile ilgili önemli iki kriter vardır. Birincisi güvenilirliği doğrulanmış veridir ikincisi ise yeterli sayıda örneklem verisidir. Bu iki kriterin ağacın güvenilirliği konusunda önemli bir yeri vardır.

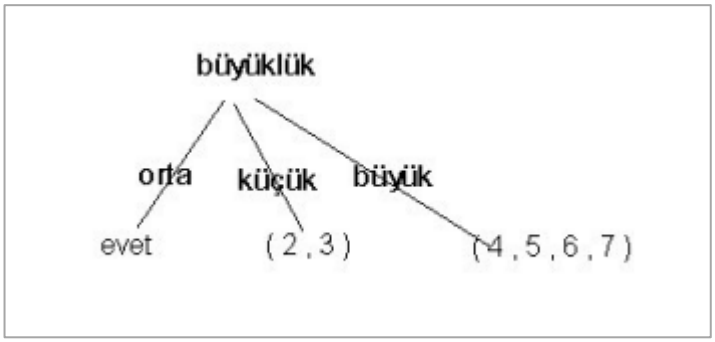
5.2.1. Böl ve Elde Et

En temel ağaç karar ağacı oluşturma metodudur. Örnek olarak bir maddenin bizim için uygun olup olmadığına bakan bir çalışma ele alınsın. Maddenin büyüklük, renk ve biçim gibi özellikleri olsun ve 7 adet de örnek olayımız olsun. Bu örnekler hayır-evet olarak ikili şekilde sınıflandırılmış olarak Şekil 5.1 ‘de gösterilmiştir.

	Büyüklük	Renk	Biçim	Sonuç
1	Orta	Mavi	Tuğla	Evet
2	Küçük	Kırmızı	Kama	Hayır
3	Küçük	Kırmızı	Küre	Evet
4	Büyük	Kırmızı	Kama	Hayır
5	Büyük	Yeşil	Sütun	Evet
6	Büyük	Kırmızı	Sütun	Hayır
7	Büyük	Yeşil	Küre	Evet

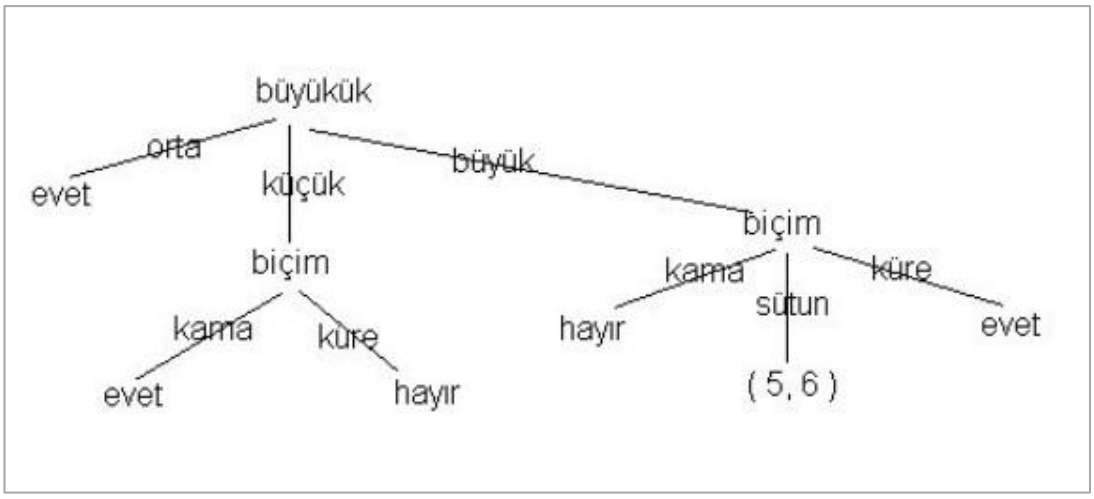
Şekil 5.1. Örnek Olay Kümesi

1'den 7'ye kadar sıralanmış örnekler seçilen büyüklük özelliğine göre alt kümelere bölünmüş olsun. Şekil 5.2'de de gösterildiği gibi büyüklüğün muhtemel üç çeşit değeri olur ve üç adet de dal oluşur.



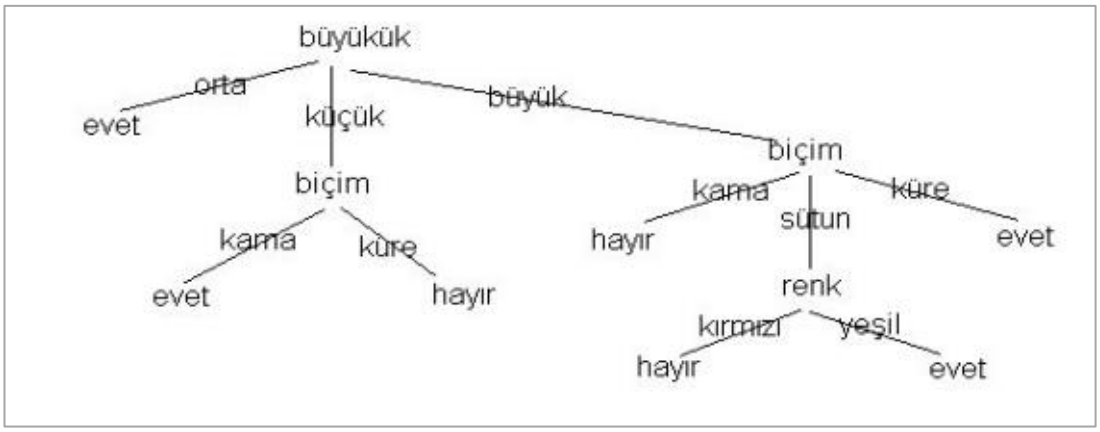
Şekil 5.2. Örnek Olay Kümesinin Büyüklük Sınıfına Göre Bölünmesi

Bu noktada küçük=büyüklük dalına ve büyük =büyüklük dalına yönelik olarak aynı işlem gerçekleştirilsin. Bölme işlemi yine rasgele seçilen biçim özelliğine göre yapıldığı takdirde Şekil 5.3 'deki ağaç oluşmuş olur.



Şekil 5.3. Ağacın Bölünmüş Kümeleri Tekrar Biçim Özelliğine Göre Tekrar Bölünmesi Sonucunda Ortaya Çıkan Ağaç

Karar Ağacında sınıflandırılmamış iki olay kaldı. Bu yüzden bölümleme işlemine devam ediliyor. Bu aşamada yine rasgele olarak renk özelliğini seçilirse Şekil 5.4 'deki karar ağacı oluşacaktır.



Şekil 5.4. Karar Ağacının Sınıflandırılmamış Olaylarının Renk Özelliği ile Sınıflandırılması

Bu aşamaya kadar tüm olaylar sınıflandırılmış oldu. Aşamaların her birinde yalnızca bir olay kaldığı için sonlandırıldı. Ancak, sadece bir tip sınıfa dair olaylar kalmış olsaydı bu durumda da sonlandırma işlemi gerçekleştirilecekti (Yıldırım, 2003, s. 9-12).

5.2.2. ID 3 Algoritması

Ross Quinlan tarafından geliştirilen bu algoritma entropi kavramından yararlanarak değişkenler içeresindeki en ayırt edici özellikteki değişkeni bulmaktadır.

Elimizde olan bilginin sayısal hale getirilmesi olarak tanımlanan entropi, veri kümesi içerisindeki rastgeleliği ve belirsizliği ölçümlemek için kullanılmaktadır. 0 ve 1 arasında bir değer alan entropi, ihtimallerin tümü eşit olduğunda ulaşabileceği en yüksek değere ulaşmış olur.

Entropi matematiksel olarak aşağıdaki şekilde ifade edilmektedir.

$$H(p1, p2 \ldots\ldots, pn) = \sum_{i=1}^{n} \pi \log(1/\pi) \quad (5.1)$$

İlk olarak veri tabanına kaydedilmiş verilerin tümü ele alınarak, veri tabanının tamamına ait bir entropisi hesaplaması yapılır (Denk 5.1). Eğer veri tabanı alt bölümlere de ayrılıyorsa, daha sonra da bu alt bölümler için de teker teker bir entropi hesaplaması yapılır. Veri tabanının entropisi bulunduktan sonra, ağaç yapısının kök kısmı ve dalları bulunur. Veri tabanının tamamı için bulunan entropi değeriyle, veriler içindeki her bir farklı değişken için bulunan değerler, ayrı ayrı çıkarılır. Bulunan her bir değer ise kazanım ölçütü olarak adlandırılır (Denk 5.2).

$$Kazanım(D, S) = H(D) - \sum_{i=1}^{n} P(Di)H(Di) \quad (5.2)$$

Hesaplanan kazanım ölçütlerinin içerisinden en büyük kısım ağacın kökü olarak belirlenir. Tekrar aynı denkleme göre (Denk 5.2), ağacın dalları bulunur. Böylece ağacın yapısı oluşturulmuş olur (EKİM, 2011, s. 23-24).

5.2.3. C4.5 Algoritması

ID3 algoritmasında birtakım sorunlar ve eksiklikler vardı. Sorunlar yine Ross Quinlan tarafından C4.5 algoritmasının geliştirilmesiyle giderilmiştir. C4.5 algoritması, ID3 algoritmasına ait tüm özelliklerini sahiplenerek oluşturulmuş bir algoritmadır. ID3 algoritmasının üzerine yeni geliştirmeler eklemiştir. Bölünme – Dağılma Bilgisi (Split - Info), özelliklerin kayıp değerlerle ile işlemler yapılması ve sayısal özellik değerlerinin hesaplanması başlıca yeniliklerdendir (Koçtürk, 2010, s. 29).

5.2.3.1. Bölünme ve Dallanma bilgisi

Sınıflandırılmış bir özelliğin olası değer çeşitliliği ne kadar yüksek olursa o özelliğin bilgi kazancı işe yaramayacak şekilde yüksek çıkar ve bu durum ağacın tamamının doğruluğunu olumsuz yönde etkiler. Bu tip özellikler işe yaramamakla

birlikte bilgi kazancı yüksek özelliklerin de önüne geçip veride gizlenmiş bağıntıların çıkarılamamasına sebep olur. Bu tip işe yaramayan bilgileri önlemek için için Ross Quinlan bölünme bilgisi kavramı ile algoritmasını güncellemiştir. Bu algoritma değer çeşitliliği fazla olan özelliklerin bilgi kazancını azaltarak algoritmanın gereksiz kestirimler yapmasının önüne geçmektedir. A bir özellik, Ai bu özelliğin değerleri, Ti, Ai özelliğinin bu veride kaç defa tekrarladığı ve T ise ele alınan olay sayısını ifade ediyorsa, bölünme bilgisi

$$-\sum_{i=1}^{k} \frac{|Ti|}{|T|} \log_2 \left(\frac{|Ti|}{T}\right) \tag{5.3}$$

şeklinde hesaplanır. Bu bölünme bilgisi tüm özelliklerin bilgi kazanç formülüne bölen olarak eklenerek bu sonuç kazanç oranı olarak ifade edilir. Mevcut durumda A özelliğinin kazanç oranı,

$$\text{Kazanç\_Oranı} = \text{Kazanım(A)} \,/\, \text{Bölünme\_Bilgisi(A)} \tag{5.4}$$

şeklinde hesaplanır (Koçtürk, 2010, s. 29).

5.2.3.2. Sayısal özellikler

Veri kümelerinde Ordinal (sayısal) ve Nominal (kategorik) olmak üzere iki tip veri vardır. ID3 algoritması yalnızca nominal tipte değerlere sahip veriler üzerinde işlem yapabiliyordu, C4.5 algoritması ile sayısal değerler üzerinde de işlem yapılabilmesi amaçlandı. Yapılması gereken eşik değerinin belirlenmesiydi. Eşik değerini belirlemek için ilk olarak sayısal ifadeler küçükten büyüğe sıralanır ve $\{v_{1,}v_2, \ldots., v_n\}$ şeklinde yazılır ve bu sıralama nitelik değerleri olarak adlandırılır. Eşik değeri, $[v_i, v_{i+1}]$ ile aralığın orta noktası olarak seçilebilir, seçilen eşik değeri ile nitelik değeri 2 parçaya ayrılır. Seçim işlemi sonrasında bütün eşik değerleri

$$t_i = \frac{v_i + v_{i+1}}{2} \tag{5.5}$$

formülü ile hesaplanabilir. Bu yapı ile söz konusu özellik küçük -büyük değerleri olan nominal bir özellik haline gelir, ardından nominal tipte değerlere uygulanan kazanım formülü tüm eşik değerler üzerinde uygulanır ve kazanımı en yüksek olan eşik değeri söz konusu özelliğin eşiği olarak kabul edilir (Yıldırım, 2003, s. 18-19).

5.2.3.3. Kayıp veriler

Veri seti içerisinde eksik verileri bulunması durumunda algoritmanın tamamen çalışması engellenir ve geçerliliği olmayan kestirimlerde bulunmasına yol açar ve bu da bir veri madenciliği uygulaması için sorun teşkil eder. Veri, farklı nedenlerden dolayı eksik olabilir, doğru formatta girilmeyen veriler olmuş olabilir, bilgi kaynağından veri tabanına aktarım sorasında hata oluşmuş olabilir vb. birden fazla sebepten dolayı veri bütünlüğü sağlanamayabilir. Bu adımda eksik bilgiler tekrar girilerek doldurulması mümkün olmayacağı için kayıp verilere ile ilgili genel yaklaşım kayıp verilerin bir şekilde tamamlanmasına yöneliktir. Kullanılan yöntemlerden biri tüm kayıp verilere aynı bilgiyi girmektir, örnek verecek olursak cinsiyet alanına boş değerler için "B" değerini girmektir. Ancak bu durumda "B" anlamlı bir sonuçmuş gibi bir netice çıkarabilir ve çok sık kullanılan bir yöntem değildir. Diğer bir yöntem ise dolu değerlerin ortalaması alınarak elde edilen değer boş değerlere yazılabilir. En sık kullanılan yöntem ise verideki kayıplar belirlenir ve ardından özelliğin kayıp değeri yerine o özelliğin en sık tekrarlanan değeri konur (Yıldırım, 2003, s. 19-20).

5.3 Ağacın Testi

Oluşturulmuş karar ağacının performansı hata oranı ölçülerek belirlenir. Karar ağaçları, her olay için bir adet sınıf tahmininde bulunur, bu tahmin doğru ise söz konusu ağaç başarılıdır değil ise başarısızdır. Hata oranı ise tüm veri kümesindeki hatalı sınıflandırılmış olayların tüm olaylara oranıdır. Bu sonuç o sınıflandırıcının performansını gösterir. Eğer bir karar ağacında budama veya kestirim işlemi gerçekleştirilmediyse genellikle o karar ağacının hata oranı %0 civarlarındadır. Bu sebeple karar ağacının oluşturulmasında kullanılan veriler ile gerçekleştirilen hata oranı ileride modele girecek olan yeni verilere ilişkin hata oranını hakkında bilgilendirici olmayacaktır. Dolayısı ile hata oranı tespiti başka, bağımsız bir veri kümesinin üzerinde yapılmalıdır. Karar ağacını yaratan verilerdeki hata oranı, genelde o karar ağacı yaratılırken kullanılan yer değiştirme işlemlerini ifade etmektedir, bu hata oranın ise geçerli olduğu söylenemez. Önemli nokta verilerden bağımsız başka veriler üzerinden bu testi uygulamaktır. Bunun için iki bağımsız veri kümesi oluşturulur.

1. Eğitim Kümesi: Karar ağacını oluşturan veri kümesi.
2. Test Kümesi: Karar ağacının hata oranını tespit eden veri kümesi.

Kısaca bir karar ağacı oluşturulmadan önce veri kümesi eğitim kümesi ve test kümesi olarak ikiye ayrılır. Bu konuda en önemli sorun elimizde ikiye bölünecek sayıda verinin olmamasıdır. Veri çok olduğunda bu işlem basitçe gerçekleştirilebilir ancak veri sayısı az ise, test için kullanacağımız veri kümesini küçültmek gerekir.
Test aşamasında seçilen bu iki küme birbirine tamamen zıt olmamalıdır. Bununla birden fazla yöntem geliştirilmiştir. Yöntemlerden biri olan ikili çapraz sağlama testi (cross validation) veri kümesini eşit iki parçaya böler; ilk olarak birini sonra diğerini eğitim kümesi olarak belirler ve bu işlemi aynı şekilde test kümesi için de gerçekleştirir. Bu durum sonucunda ortaya çıkmış olan ortalama hata değer daha gerçekçi bir sonuç verecektir. Ancak bazı çalışmalar da bu sonucun da yetersiz kaldığı gözlemlenmiştir. Gözlemler sonucunda ikili sağlama işleminin yaptığı ikiye bölme yerine ona bölünerek test edilmiştir. Onlu çapraz sağlama işlemi ilk olarak verinin tamamını on eşit parçaya bölümler. Ardından her birini sırayla dışarı alır ve geriye kalmış olan dokuz adet veri kümesinden bir eğitim seti oluşturarak karar ağacını oluşturur. Ardından bu dışarıya aldığı veri kümesi ile test eder. Bu işlemi her biri için tekrar tekrar yapar ve işlemlerden çıkan on adet hata oranının ortalamasını geçerli ortalama olarak kabul eder (Koçtürk, 2010, s. 30-31).

5.4. Ağacın Budanması

Veri yığınlarından oluşan karar ağacının çok fazla dalı olabilir. Örnekleyecek olursak dal sayısı 15 olan bir karar ağacında 15 adet if (eğer) koşulundan sonra bir neticeye ulaşılıyor demektir. Veya bir duruma 15 adet soru sorduktan sonra o olay ile ilgili konuşulabiliyor demektir. Bu pek karşılaşılmak istenmeyen bir şeydir, bir karar ağacında dal sayısının fazla olması doğruluk oranının düşmesine sebep olur. Basit bir veri yığınında karar ağacının çok büyük çıkması şişme (overfitting)olarak adlandırılır. Böyle bir sonuç çıkmasının iki temel vardır. İlki veri yığını içerisinde gürültülü veri bulunmaktadır, veri temizleme aşaması atlanmıştır. Gürültü veri yığınının geçerliliği olmayan, anlam ifade etmeyen kurallar türetmesine yol açar. Bu durumda karar ağacından oluşturulan kurallar gürültülü veri ile ilgilidir. Diğer ihtimal ise seçilen veri kümesinin o olayı temsil etmemesidir. Bu durumlarda karar ağacı budanmalıdır.

Bazı durumlarda örneklem kümesini daha fazla bölmemek istenebilir. Bölme işlemine son verme kararı ki-kare gibi istatistiksel testler ile alınabilir. Bölünme

öncesine ve sonrasında önemli bir fark yoksa o zaman geçerli düğüm bir yaprak olarak gösterilir. Bu da bir ön budamaya örnektir.

Seçilen bir doğruluk ölçütü kullanarak bazı karar ağaçları budanabilir. Bu yöntem karar ağacı oluşturulduktan sonra uygulanır. Bundan sebepten bu budama yöntemine sonradan budama olarak adlandırılmıştır. Bu durumda hata oranlarına göre alt ağaç, kök düğüm haline getirilerek budama işlemi tamamlanmış olur (Gemici, 2012, s. 34-36).

6. UYGULAMA

Başlık altında veri madenciliği adımları gerçekleştirilerek ve açıklanarak bir veri madenciliği uygulaması geliştirilecektir.

6.1. Problemin Tanımlanması Aşaması

Sigorta sektöründe çok sayıda kendilerine haksız menfaat sağlamak isteyen kişiler görülmektedir, uygulama içersinde de bir sigorta şirketinin bir müşteriye araç sigorta poliçesi düzenlemeden önce 1993 ve 1995 yılları arasında toplanmış 15421 satır veriden oluşan araç sigorta poliçe dolandırıcılığına ilişkin verileri inceleyerek kendilerine haksız menfaat sağlayabilecek müşterileri belirleyerek o müşterilerin elenmesinde hangi değişkenlerin en önemli olduğunu belirleyerek sigorta şirketlerine yardımcı olacak bir veri madenciliği uygulaması yapılacaktır.

6.2. Veri Tanıma Aşaması

Verilerimiz 1993 ve 1995 yılları arasında araç sigorta poliçesi dolandırıcılığına ilişkin verileridir. 33 sütun 15419 satırdan oluşmaktadır. Veri bilimi platrformu Kaggle üzerinden veri setine erişilmiştir.

- FraudFound_P: Karar değişkenidir, ilgili satırda dolandırıcılık olup olmadığını konusunda bilgi verir.
- Month: Kazanın gerçekleştiği ay hakkında bilgi verir.
- DayOfWeek: Kazanın gerçekleştiği gün hakkında bilgi verir
- Make: Kazaya karışan aracın markası hakkında bilgi verir.
- AccidentArea: Kazanın gerçekleştiği bölge hakkında bilgi verir.
- Sex: Poliçe sahibinin cinsiyetinin bildirir.
- MaritalStatus: Poliçe düzenlenen kişinin medeni durumu hakkında bilgi verir.
- Age:Poliçe düzenlenen kişinin yaşı hakkında bilgi verir.

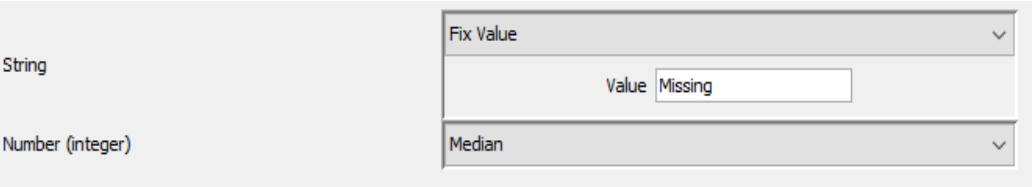
- Fault:Kazada hatanın kimde olduğu konusunda bilgi verir, poliçe sahibi veya diğer taraf.
- PolicyType: Poliçenin düzenlendiği araç ve poliçe hakkında bilgi verir.
- VehicleCategory: Yalnıza araç tipi hakkında bilgi verir.
- VehiclePrice: Poliçe düzenlenen aracın fiyatını belirli aralıklarda sınıflandırarak aracın fiyatı hakkında bilgi verir.
- PolicyNumber: Tabloda tekliği sağlamak için her bir poliçe numaralandırılmıştır.
- RepNumber: Poliçe sahibinin aynı firmadan kaç kere poliçe yenildiğini bildirir
- DriverRating: Poliçe sahipleri 1 ile 5 arasında seviyelendirilmiştir, bu konuda bilgi verir.
- PastNumberOfClaims: Poliçe sahibinin geçmişte yaptığı kaza sayısı hakkında bilgi verir.
- AgeOfVehicle: Poliçe düzenlenen aracın yaşı hakkında bilgi verir.
- AgeOfPolicyHolder: Poliçe sahibinin yaşı hakkında bilgi verir, belirli aralıklarda sınıflandırılmıştır.
- PoliceReportFiled: Kazanda sonra polis tarafında bir tutanak tutulup tutulmadığı hakkında bilgi verir.
- WitnessPresent: Kaza sırasında şahit olup olmadığı hakkında bilgi verir.
- AddressChange_Claim: Poliçe sahibinin adres değişikliği yapıp yapmadığı veya kaç kere yaptığı hakkında bilgi verir.
- NumberOfCars: Kazaya kaç aracın karıştığını belirtir.
- BasePolicy: Yalnızca poliçe tipi hakkında bilgi veri.
- Year: Kaza yılı hakkında bilgi verir.

6.3. Veri Hazırlama Aşaması

- AgeOfPolicyHolder değişkeni üzerinde yaş bilgisi belirli aralıklar üzerinden verildiği ve değerlendirmelerde modele daha çok katkı sağladığı görüldüğü için age değişkeni filtrelendi.
- PolicyType değişkeni poliçe tipi ve araç tipi hakkında tek bir sütunda bilgi verir, bu bilgileri BasePolicy ve VehicleCategory değişkenleri ayrı ayrı

sütunlarda verdiği için ve veri tekrarı yapmamak için PolicyType değişkeni filtrelendi.

- PolicyNumber değişkeni veriler arasındaki tekliği kontrol ettiği başka bir poliçede tekrar aynı numara olamayacağı için filtrelendi.
- Year değişkeni geçmişte kaldığı, tekrarlanamayacağı için filtrelendi.
- Deductible, Days_Policy_Accident, Days_Policy_Claim, NumberOfSuppliments, değişkenleri modele katkı vermediği için filtrelendi
- Kullanılan veri seti içerisinde eksik veri bulunmamasına rağmen veri madenciliği aşamalarını tam olarak uygulamak için sözel değişkenlerde kayıp veriler için boş olan tüm değerlere "Missing" olarak yeni bir değer atandı. Sayısal değişkenlerde boş değerleri aynı değişkenlerin ortalamasını alarak boş olan değere yazılmasını sağlandı.



Şekil 6.1. Kayıp Veriler Üzerinde Yapılan İşlemler

- Karar değişkenimiz, FraudFound_P üzerinde değerlerin daha anlaşılır olabilmesi için değer 1 ise "There is fraud" dolandırıcılık var, değer 0 ise "There is not fraud" dolandırıcılık yok şeklinde değiştirildi.

```
$FraudFound_P$= 1 => "There is fraud"
$FraudFound_P$= 0 => "There is not fraud"
```

Şekil 6.2. Karar Değişkeni Üzerinde İşlem

- Göz ile daha iyi ayırt edebilmek için karar değişkenimiz olan FraudFound_P üzerinde There is fraud olanları yeşil, There is not fraud olanları kırmızı olarak renklendirildi.

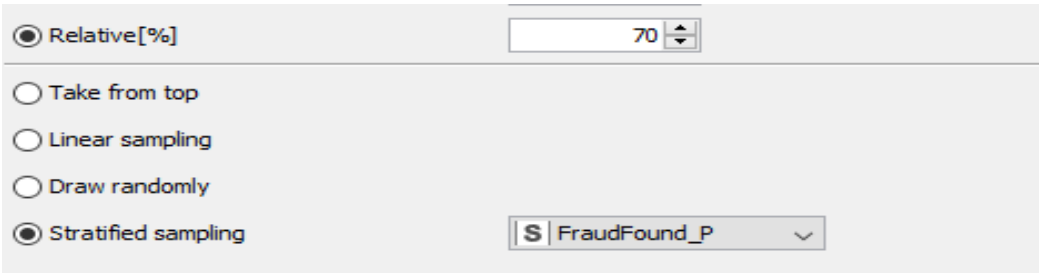
Row49	Apr	3	Friday	Pontiac	Urban	Tuesday	Apr	3	Male	Married	Policy Holder	Sedan	20000 to 29000	0	4
Row50	May	4	Friday	Pontiac	Urban	Friday	May	4	Male	Married	Policy Holder	Sport	30000 to 39000	0	2
Row51	Jun	4	Wednesday	Pontiac	Urban	Friday	Jun	4	Male	Married	Policy Holder	Sport	20000 to 29000	0	3
Row52	Jul	3	Sunday	Honda	Rural	Wednesday	Jan	4	Male	Married	Policy Holder	Sport	more than 69000	1	4
Row53	Jul	4	Saturday	Honda	Urban	Wednesday	Aug	2	Male	Married	Policy Holder	Sedan	20000 to 29000	1	1

Şekil 6.3. Renklendirme İşlemi Yapılmadan Önce

Row50	May	4	Friday	Pontiac	Urban	Friday	May	4	Male	Married	Policy Holder	Sport	30000 to 39000	There is not fraud
Row51	Jun	4	Wednesday	Pontiac	Urban	Friday	Jun	4	Male	Married	Policy Holder	Sport	20000 to 29000	There is not fraud
Row52	Jul	3	Sunday	Honda	Rural	Wednesday	Jan	4	Male	Married	Policy Holder	Sport	more than 69000	There is fraud
Row53	Jul	4	Saturday	Honda	Urban	Wednesday	Aug	2	Male	Married	Policy Holder	Sedan	20000 to 29000	There is fraud
Row54	Jun	4	Tuesday	Pontiac	Urban	Tuesday	Jun	4	Male	Single	Policy Holder	Sedan	30000 to 39000	There is not fraud
Row55	Dec	3	Monday	Mazda	Urban	Wednesday	Dec	3	Male	Married	Third Party	Sedan	20000 to 29000	There is not fraud

Şekil 6.4. Renklendirme İşlemi Yapıldıktan Sonra

- Verimizi %70 eğitim %30 test verisi olmak üzere ikiye böldük. Bölünmüş her iki alanda da karar değişkenimiz ona FraudFound'ın eşit oranda olması için "Stratified sampling" alanını işaretlendi.

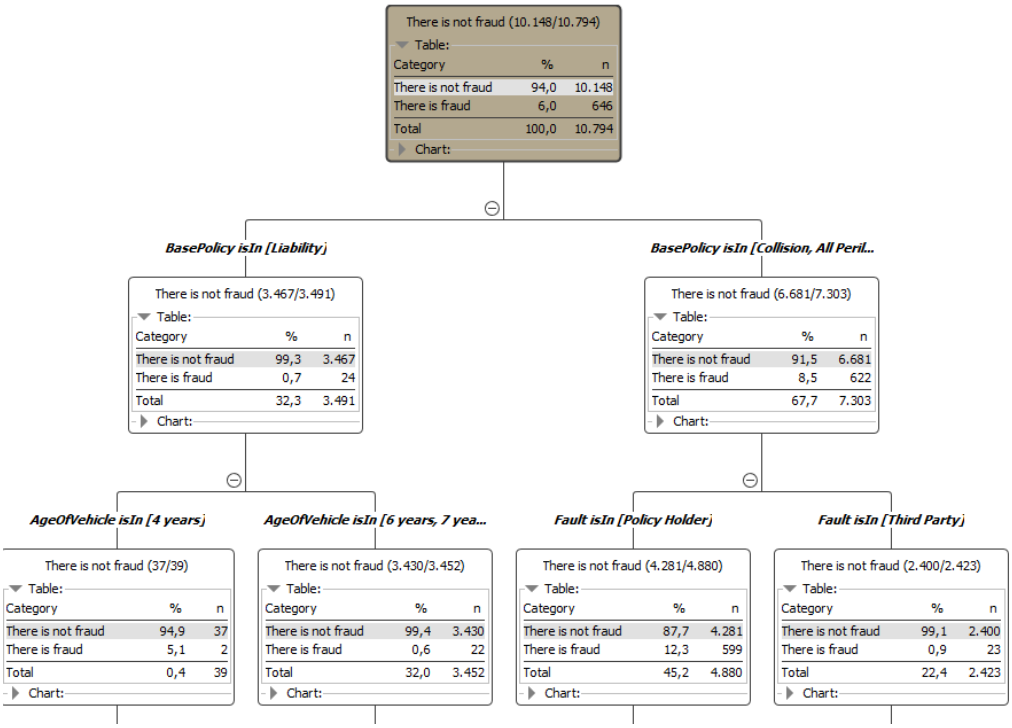


Şekil 6.5. Verinin Bölünmesi

6.4. Modelleme Aşaması

Veri seti üzerinde sonuçlandırılmış olaylar referans alınarak ileride karşılaşabileceğimiz benzer durumlarda karar vermemize yardımcı olacak bir karar ağacı oluşturulmuştur.

Karar ağacımızın ilk 3 satırı aşağıdaki şekilde oluşmuştur



Şekil 6.6. Karar Ağacı

Karar ağacımıza göre oluşan 533 adet karar kuralından bazıları aşağıda listelenmektedir.

- $DayOfWeek$ IN ("Tuesday", "Sunday") AND $Make$ IN ("Toyota") AND $AgeOfVehicle$ IN ("4 years") AND $BasePolicy$ IN ("Liability") => "There is fraud"
- $Make$ IN ("Mercury") AND $AgeOfVehicle$ IN ("6 years", "7 years", "more than 7", "5 years", "new", "3 years", "2 years") AND $BasePolicy$ IN ("Liability") => "There is not fraud"
- $DayOfWeek$ IN ("Sunday", "Monday") AND $DayOfWeekClaimed$ IN ("Saturday") AND $Fault$ IN ("Third Party") AND $BasePolicy$ IN ("Collision", "All Perils") => "There is fraud"
- $DayOfWeek$ IN ("Wednesday", "Friday", "Saturday", "Tuesday", "Thursday") AND $DayOfWeekClaimed$ IN ("Saturday") AND $Fault$ IN ("Third Party") AND $BasePolicy$ IN ("Collision", "All Perils") => "There is not fraud"
- $Make$ IN ("Honda", "Ford", "Mazda", "Chevrolet", "Pontiac", "Accura", "Dodge", "Mercury", "Jaguar", "Saab", "VW", "Saturn", "Porche", "Nisson", "Mecedes", "BMW", "Ferrari", "Lexus") AND $AgeOfVehicle$ IN ("4 years") AND $BasePolicy$ IN ("Liability") => "There is not fraud"

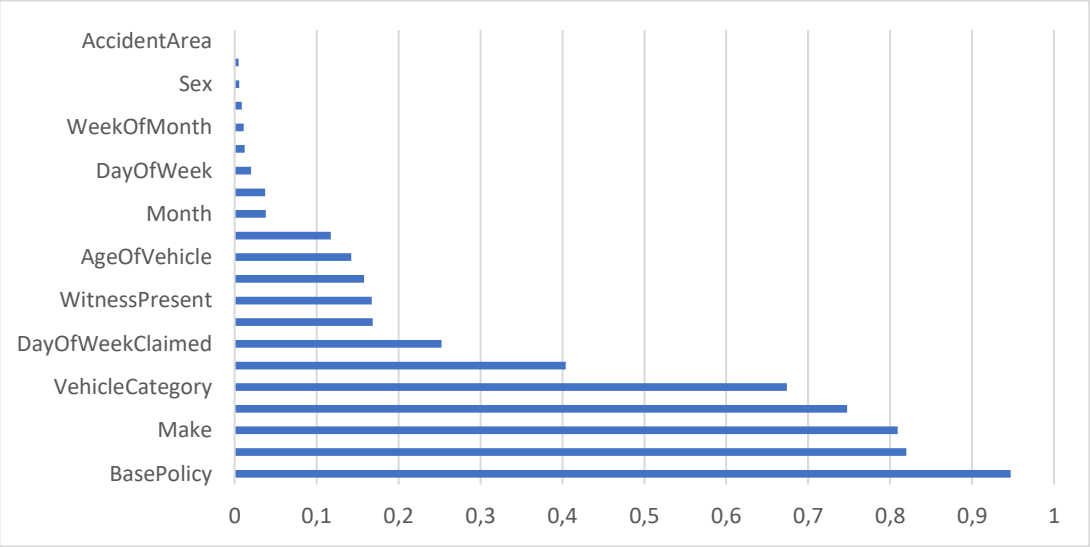
- $DayOfWeek$ IN ("Wednesday", "Friday", "Saturday", "Thursday", "Monday") AND $Make$ IN ("Toyota") AND $AgeOfVehicle$ IN ("4 years") AND $BasePolicy$ IN ("Liability") => "There is not fraud"
- $MonthClaimed$ IN ("Nov", "Jul", "Dec", "Apr", "Mar", "Feb", "Aug", "Jun", "Sep", "Oct", "0") AND $MonthClaimed$ IN ("Jan", "Nov", "Jul", "Dec", "Apr", "Feb", "Aug", "May", "Jun", "Sep", "Oct", "0") AND $WeekOfMonth$ > 1.5 AND $Make$ IN ("Toyota", "Chevrolet", "Pontiac", "Jaguar", "VW", "Saturn", "Nisson") AND $ï»¿Month$ IN ("Jan", "Oct", "Jun", "Apr", "Jul", "May", "Nov", "Feb", "Aug") AND $PastNumberOfClaims$ IN ("none", "more than 4") AND $DayOfWeek$ IN ("Saturday", "Tuesday", "Monday") AND $AgeOfPolicyHolder$ IN ("41 to 50", "51 to 65", "36 to 40", "over 65", "26 to 30", "18 to 20") AND $BasePolicy$ IN ("Liability", "All Perils") AND $MonthClaimed$ IN ("Jan", "Apr", "Mar", "Feb", "Aug", "May", "Jun", "Sep", "Oct") AND $Fault$ IN ("Policy Holder") AND $BasePolicy$ IN ("Collision", "All Perils") => "There is not fraud"
- $PastNumberOfClaims$ IN ("1", "2 to 4", "more than 4") AND $VehiclePrice$ IN ("20000 to 29000") AND $MonthClaimed$ IN ("Apr", "Jun") AND $DayOfWeekClaimed$ IN ("Friday") AND $ï»¿Month$ IN ("Apr", "May", "Aug") AND $VehiclePrice$ IN ("more than 69000", "20000 to 29000", "30000 to 39000", "40000 to 59000", "60000 to 69000") AND $ï»¿Month$ IN ("Oct", "Jun", "Apr", "Mar", "Jul", "May", "Sep", "Aug") AND $VehiclePrice$ IN ("more than 69000", "20000 to 29000", "30000 to 39000", "less than 20000", "40000 to 59000") AND $AgeOfPolicyHolder$ IN ("31 to 35", "36 to 40") AND $DayOfWeekClaimed$ IN ("Monday", "Thursday", "Friday", "Wednesday", "Tuesday", "Sunday", "0") AND $Fault$ IN ("Third Party") AND $BasePolicy$ IN ("Collision", "All Perils") => "There is not fraud"
- $AgeOfVehicle$ IN ("4 years") AND $Make$ IN ("Honda", "Toyota", "Mazda", "Chevrolet", "Pontiac", "Accura", "Dodge", "Mercury", "Jaguar", "Saab", "Saturn", "Porche", "Nisson", "Mecedes", "BMW", "Ferrari", "Lexus") AND $WeekOfMonthClaimed$ <= 4.5 AND $ï»¿Month$ IN ("Jan", "Dec", "Nov", "Feb") AND $VehiclePrice$ IN

("more than 69000", "20000 to 29000", "30000 to 39000", "less than 20000", "40000 to 59000") AND $AgeOfPolicyHolder$ IN ("31 to 35", "36 to 40") AND $DayOfWeekClaimed$ IN ("Monday", "Thursday", "Friday", "Wednesday", "Tuesday", "Sunday", "0") AND $Fault$ IN ("Third Party") AND $BasePolicy$ IN ("Collision", "All Perils") => "There is not fraud"

- $AgeOfVehicle$ IN ("7 years") AND $MonthClaimed$ IN ("Jan", "Nov", "Jul", "Dec", "Apr", "Aug", "May", "Sep", "Oct", "0") AND $WeekOfMonth$ > 3.5 AND $Make$ IN ("Honda", "Ford", "Mazda", "Accura", "Dodge", "Mercury", "Saab", "Porche", "Mecedes", "BMW", "Ferrari", "Lexus") AND $ï»¿Month$ IN ("Jan", "Oct", "Jun", "Apr", "Jul", "May", "Nov", "Feb", "Aug") AND $PastNumberOfClaims$ IN ("none", "more than 4") AND $DayOfWeek$ IN ("Saturday", "Tuesday", "Monday") AND $AgeOfPolicyHolder$ IN ("41 to 50", "51 to 65", "36 to 40", "over 65", "26 to 30", "18 to 20") AND $BasePolicy$ IN ("Liability", "All Perils") AND $MonthClaimed$ IN ("Jan", "Apr", "Mar", "Feb", "Aug", "May", "Jun", "Sep", "Oct") AND $Fault$ IN ("Policy Holder") AND $BasePolicy$ IN ("Collision", "All Perils") => "There is not fraud"
- $MonthClaimed$ IN ("Nov", "Jul", "Apr", "Mar", "Aug", "May", "Jun", "Sep", "Oct", "0") AND $Make$ IN ("Ford", "VW") AND $WeekOfMonthClaimed$ <= 4.5 AND $ï»¿Month$ IN ("Jan", "Dec", "Nov", "Feb") AND $VehiclePrice$ IN ("more than 69000", "20000 to 29000", "30000 to 39000", "less than 20000", "40000 to 59000") AND $AgeOfPolicyHolder$ IN ("31 to 35", "36 to 40") AND $DayOfWeekClaimed$ IN ("Monday", "Thursday", "Friday", "Wednesday", "Tuesday", "Sunday", "0") AND $Fault$ IN ("Third Party") AND $BasePolicy$ IN ("Collision", "All Perils") => "There is fraud"
- $MonthClaimed$ IN ("Nov", "Jul", "Apr", "Mar", "Aug", "May", "Jun", "Sep", "Oct", "0") AND $Make$ IN ("Ford", "VW") AND $WeekOfMonthClaimed$ <= 4.5 AND $ï»¿Month$ IN ("Jan", "Dec", "Nov", "Feb") AND $VehiclePrice$ IN ("more than 69000", "20000 to 29000", "30000 to 39000", "less than 20000", "40000 to 59000") AND $Fault$ IN ("Third Party") AND $BasePolicy$ IN ("Collision", "All Perils") => "There is fraud"

6.5. Değerlendirme Aşaması

Karar vermemizde en çok yardımcı olan değişkenler aşağıda gösterilmektedir



Şekil 6.7. Karara En Çok Etki Eden Değişkenler

Modelimiz eğitim verilerinden aldığı bilgileri test verileri üzerinde kullanarak %88,716 doğruluk oranı sağlamıştır. Dolandırıcılık olmayan 4065 kaydı doğru tahmin edip 238 tanesinde ise yanılmıştır. Dolandırıcılık olanlarda ise 39 tanesini doğru tahmin edip 284 tanesinde yanılmıştır.

FraudFound_P \ Prediction (FraudFound_P)	There is not fraud	There is fraud
There is not fraud	4065	284
There is fraud	238	39
Correct classified: 4.104		Wrong classified: 522
Accuracy: 88,716 %		Error: 11,284 %

Şekil 6.8. Hata Oranı

6.6. Uygulama Aşaması

Uygulamamızı kullanarak bir sigorta şirketi elde edilen bilgiler doğrultusunda müşterileri arasında seçim yapabilir, düzenleneceği poliçede aldığı riski daha önceden öngörebildiği için poliçe bedelinde değişiklik yapabilir, şüpheli durumlarda oluşturulan model ile karşı karşıya kaldığı durum arasında bir kıyas yaparak bir fikir sahibi olabilir, uygulamamızı bir karar destek sistemi olarak kullanabilir.

7. SONUÇLAR

Elimizde bulunan verinin miktarı arttıkça veriden bilgi elde etmek daha zor bir hale gelecektir. Bu aşamada da veri madenciliği devreye girer ve yüksek hacimli verilerde geleceğe yönelik tahminler yapmamıza olanak sağlayarak içerisinde bulunduğumuz durumlarda henüz durum sonlandırılmadan bir öngörüde bulunma olanağı sağlar. Veri madenciliği günümüzde birçok alanda kullanılabilir. Örneğin sosyal medyada duygu analizi, otonom araçlarda araçların karşılaştığı durumlarda vereceği tepkiler veri madenciliği sayesinde yapılabilmektedir. İşletmelerde ise veri madenciliği sayesinde firmalar müşterilerini tanıyarak müşterilerine özel kampanyalar yaparak satışlarını arttırabilirler, bulundukları pazarın analizini yaparak gereksiz hammadde stoğu yapmanın önüne geçebilir. Direkt olarak satılacak ürün stoklarını belirleyerek nakit akışlarını düzenleyebilirler. Sepet analizi yaparak tüketicilerin hangi ürünleri beraber aldıklarını belirleyerek fiziksel mağazada raf düzenini belirleyerek müşterinin o an için aklında olmayan bir ürünü müşterinin satın almasını sağlayabilirler. Online satış sitelerinde ürün önerisi yaparak satışlarını arttırabilir. Yapılmış olan uygulamada olduğu gibi satış sürecinden önce müşteri veya pazar üzerinde bir araştırma yaparak ürün satışı konusunda çekimser kalabilir. Alacağı risk neticesinde fiyat üzerinde değişiklik yapabilir. Bu sayılan nedenler ile firmalar hangi sektörde olursa olsun oluşan rekabette rakip firmalara göre avantaj sağlamak için veri madenciliği uygulamalarını kullanmaları gerekmektedir.

Yapılan araştırmalarda ise sigortacılık sektöründe birçok dolandırıcılık örneği paylaşılmıştır. Sigortalı tarafından yapılabilecek bir dolandırıcılık durumunda sigorta şirketlerinin veri madenciliği kullanarak alabileceği önlemlerden birini 1993-1995 yılları arasında Amerika'da toplanan araç sigorta poliçe dolandırıcılığı veri seti üzerinde KNIME adlı açık kaynak kodlu bir program ile çalıştık. İlk olarak değişkenlerimizin tiplerini bizim için ne ifade ettiğini belirledik ardından veri setimiz temiz bir veri seti olmasına rağmen veri madencliği aşamalarını tam olarak uygulayabilmek için değişkenlerimiz üzerinde kirli ve gürültülü verileri azaltmaya yönelik bir işlem yaptık. Verimizi %70 eğitim %30 test verisi olmak üzere ikiye

bölerek modelimizin doğruluğunu sınadık. Elde ettiğimiz bilgileri incelemeye başladık ve bir karar ağacı oluşturduk. Çalıştığımız veri setine göre bir sigortalının sahtekarlık yapıp yapmayacağını belirleyen en önemli değişken "Base Policy" olmuştur. "Base Policy" değişkeni "Liability" iken sahtekarlık yapma oranı %0,7, diğer değerlerin toplamı %8,5'tir. Oluşturulan modelin doğruluk oranı ise %88,716'dır bu değer de başarılı bir değer olarak kabul edilebilir. Oranın iyi bir seviyede çıkması kullanılan veri setinin temiz, kirli, gürültülü veri barındırmaması, iyi bir veri seti olması ile ilişkilendirilebilir.

Özetleyecek olursak artan rekabet koşullarında kendilerini diğer işletmelere göre kendilerine avantaj sağlamak zorunda olan işletimlerin veri madenciliğini etkin bir şekilde kullanarak sağlayabilecekleri avantajları göstermek istedik ve sigortacılık sektöründeki sorunlardan biri olan poliçe suiistimali ile ilgili bir veri madenciliği uygulaması ile destekledik.

Çalışma içerisinde paylaşılan bu sonuçların literatüre katkıda bulunacağını umut ediyorum.

KAYNAKLAR

Akpınar, Ö. (2018). Sigorta sektöründe veri madenciliği ve kullanım alanları. *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi.*

Akyiğit, H. E. (2021). *Sigortacılık sektöründe makine öğrenmesi ile müşteri kaybı analizi*. (Yayımlanmamış Tez). Sakarya Üniversitesi, Sakarya.

Alpaydın, E. (2000). Ham veriden altın bilgiye ulaşma yöntemleri. Bilişim 2000 Eğitim Semineri (s. 1). İştanbul.

Altun, A. (2007). *Sigortacılık Sektöründe Acentelerin Önemi T.C. Kadir Has Üniversitesi Sosyal Bilimler Enstitüsü*, Yüksek Lisans Tezi, İstambul.

Ekim, U. (2011). *Veri madenciliği algoritmalarını kullanarak öğrenci*. Konya.

Erol, B. (2013). *Müşteri ilişkileri yönetimi için veri madenciliği kullanılması ve sigortacılık sektörü üzerine bir uygulama*. (Yayımlanmamış Tez). Marmara Üniversitesi, İstanbul.

Gemici, B. (2012). *Veri madenciliği ve bir uygulaması*. (Yayımlanmamış Tez). Dokuz Eylül Üniversitesi, İzmir.

Gündüz, H. (2020). WEKA veri madenciliği yazılımının sürümleri. BŞEÜ Fen Bilimleri Dergisi arasındaki kalite değişimlerinin QMOOD ile incelenmesi.

Gürel, A. G. (2019). *Üniversite kütüphanesi verileri üzerinde veri madenciliği yöntemlerinin uygulanması.* (Yayımlanmamış Tez). Afyon Kocatepe Üniversitesi, Afyon.

Haberal, İ. (2007). *Veri madencliği algoritmaları kullanılarak web günlük erişimlerinin analizi.* (Yayımlanmamış Tez). Başkent Üniversitesi, Ankara.

Karadağ, H. (2021). *Sigortacılık sektöründe siber güvenlik yönetimi ve riskin azaltılmasında siber güvenlik sigortalarının rolü*. (Yayımlanmamış Tez). Marmara Üniversitesi, İstanbul.

Kasap, E. (2007). *Sigortacılık sektöründe müşteri ilişkileri yönetimi yaklaşımıyla verimadenciliği teknikleri ve bir uygulama*. (Yayımlanmamış Tez). Marmara Üniversitesi, İstanbul.

Kavurkacı, Ş., Gürkaş Aydın, Z. ve Şamlı, R. (2011). Büyük ölçekli veri tabanlarında bilgi keşfi. Malatya.

Koçtürk, Y. (2010). *Veri madenciliğinde bağlılık*. (Yayımlanmamış Tez). İstanbul Teknik Üniversitesi, İstanbul.

Muslu, D. (2009). *Sigortacılık sektöründe risk analizi: veri madenciliği uygulaması*. (Yayımlanmamış Tez). İstanbul Teknik Üniversitesi, İstanbul.

Sarıoğuz, S. (2007). *Sigorta sektöründe risk yönetimi, alternatif risk transfer yöntemleri, şirketler için bir öneri: Hava durumu opsiyonları*. (Yayımlanmamış Tez). İstanbul Üniversitesi, İstanbul.

Şen, F. (2008). *Veri madenciliği ile birliktelik kurallarının bulunması*. (Yayımlanmamış Tez). Sakarya Üniversitesi, Sakarya.

Uluyardımcı, M. M., ve Zontul, M. (2020). Veri madenciliği yöntemleri ile uçuş biletleme analizi. *Aurum Mühendislik Sistemleri ve Mimarlık Dergisi*, 4.

Varol, N. (2021). *Sigortacılıkta dijitalleşme: Sigorta sektöründeki müşteriler ve acente çalışanları üzerine Ankara ilinde bir uygulama*. (Yayımlanmamış Tez). Ankara Hacı Bayram Veli Üniversitesi, Ankara.

Yayla, Ş. O. (2019). Sigortacılık ve Türkiye'de sigorta sektörünün durumu. *Liberal Düşünce Dergisi*, 108-109.

Yıldırım, S. (2003). *Tümevarım öğrenme tekniklerinden C4.5'in incelenmesi*. (Yayımlanmamış Tez). İstanbul Teknik Üniversitesi, İstanbul.

Zaimoğlu, E. A. (2018). *Veri madenciliği teknikleri kullanılarak sosyal ağlar aracılığı ile bilgisayar ve bilişim mühendisliği mezun öğrenci profillerinin belirlenmesi*. (Yayımlanmamış Tez). Sakarya Üniversitesi, Sakarya.

URL-1 https://www.kaggle.com/datasets/shivamb/vehicle-claim-fraud-detection/, Erişim tarihi: 22.05.2022 adresinden erişildi

URL-2 https://web.stanford.edu/, Erişim tarihi: 22.05.2022 adresinden erişildi

ÖZGEÇMİŞ

Ad-Soyad **:** Taha Korkmaz

ÖĞRENİM DURUMU:

- **Lisans** **:** 2020, Anadolu Üniversitesi/Açıköğretim Fakültesi/Yönetim Bilişim Sistemleri
- **Yüksek Lisans** **:** Devam Ediyor, Haliç Üniversitesi/Lisansüstü Eğitim Enstitüsü/Yönetim Bilişim Sistemleri

MESLEKİ DENEYİM VE ÖDÜLLER:

Mitra Bilgisayar- Yazılım Geliştirici