T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YURT DIŞI DİL EĞİTİM KURUMLARINDA VERİ MADENCİLİĞİ YÖNTEMİ İLE HEDEF KİTLE BELİRLEME

YÜKSEK LİSANS TEZİ

Sevim Şevval ZOROĞLU

Endüstri Mühendisliği Anabilim Dalı

ŞUBAT 2023

T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YURT DIŞI DİL EĞİTİM KURUMLARINDA VERİ MADENCİLİĞİ YÖNTEMİ İLE HEDEF KİTLE BELİRLEME

YÜKSEK LİSANS TEZİ

Sevim Şevval ZOROĞLU

Endüstri Mühendisliği Anabilim Dalı

Tez Danışmanı: Doç Dr. Ayten YILMAZ YALÇINER

ŞUBAT 2023

Sevim Şevval Zoroğlu tarafından hazırlanan "Yurt Dışı Dil Eğitim Kurumlarında Veri Madenciliği Yöntemi ile Hedef Kitle Belirleme " adlı tez çalışması 15.02.2023 tarihinde aşağıdaki jüri tarafından oy birliği/oy çokluğu ile Sakarya Üniversitesi Fen Bilimleri Enstitüsü Endüstri Mühendisliği Anabilim Dalı Yüksek Lisans tezi olarak kabul edilmiştir.

Tez Jürisi

Jüri Başkanı : Doç. Dr. Tijen Över ÖZÇELİK
Sakarya Üniversitesi

Jüri Üyesi : Doç. Dr. Ayten YILMAZ YALÇINER
Sakarya Üniversitesi

Jüri Üyesi : Dr. Öğr. Üyesi Mustafa YILMAZ
Sakarya Üniversitesi

ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ

Sakarya Üniversitesi Fen Bilimleri Enstitüsü Lisansüstü Eğitim-Öğretim Yönetmeliğine ve Yükseköğretim Kurumları Bilimsel Araştırma ve Yayın Etiği Yönergesine uygun olarak hazırlamış olduğum "Yurt Dışı Dil Eğitim Kurumlarında Veri Madenciliği Yöntemi İle Hedef Kitle Belirleme" başlıklı tezin bana ait, özgün bir çalışma olduğunu; çalışmamın tüm aşamalarında yukarıda belirtilen yönetmelik ve yönergeye uygun davrandığımı, tezin içerdiği yenilik ve sonuçları başka bir yerden almadığımı, tezde kullandığım eserleri usulüne göre kaynak olarak gösterdiğimi, bu tezi başka bir bilim kuruluna akademik amaç ve unvan almak amacıyla vermediğimi ve 20.04.2016 tarihli Resmi Gazete'de yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince Sakarya Üniversitesi'nin abonesi olduğu intihal yazılım programı kullanılarak Enstitü tarafından belirlenmiş ölçütlere uygun rapor alındığını, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun ortaya çıkması halinde doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi beyan ederim.

(05/01/2023)

Sevim Şevval Zoroğlu

Kendime...

TEŞEKKÜR

Bütün eğitim hayatım boyunca yanımda olan beni destekleyen aileme teşekkürlerimi sunuyorum.

Sevim Şevval Zoroğlu

İÇİNDEKİLER

TABLO LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

YURT DIŞI DİL EĞİTİM KURUMLARINDA VERİ MADENCİLİĞİ YÖNTEMİ İLE HEDEF KİTLE BELİRLEME

ÖZET

Günümüzde teknolojinin gelişmesi ve beraberinde getirdiği yenilikler ile yabancı dil öğrenimin ve öneminin oranı ciddi olarak artmıştır. Global düzende yabancı dil öğrenmek sektör fark etmeksizin profesyonel iş yaşamında da oldukça kritik bir öneme sahip olmaya başlamıştır. Özellikle bu önem özel sektör çalışanlarında daha fazla görülmeye başlanmıştır. Özel sektörde kaynaklanan bu önemden dolayı dil bilmek işe alımlarda en belirleyici faktörlerden biri olmıuştur. Üniversite öğrencileri çoğunlukla hem maliyet hem zaman açısından öğrenci değişim programlarını tercih etse de bu durum yetersiz gelmeye başlamıştır. Bu derece öneme sahip kriter için Türkiye'de de her kesim bireyin yararlanabileceği çok fazla yurt dışı eğitim danışmanlık şirketleri kurulmuştur. Bu kuruluşlar her kesime her yaşa hitap eden hem eğitim içerikli hem çalışma içerikli hemde seyahat etmeli programlar uygulamaktadırlar. Gün geçtikçe bu kuruluşların sayısının artması kuruluşlar arasında bir rekabete yol açmaya başlamıştır. Çalışmada ele alınan danışmanlık şirketinin müşteri bilgileri toplanarak hangi kriterlerin kayıt durumunu etkilediğinin öğrenilmesi amaçlanmıştır. Bu sonuca göre müşterilerin tanımlanması, hedef müşteri grubunun belirlenmesi, özel teklif sunulabilecek müşterini profilinin belirlenmesi, gerekli durumlarda müşteri portföyünün değiştirilmesi, en çok müşteri aldıkları zamanların belirlenmesi, gerekli iyileştirme önerilerinin sunulması hedeflenmektedir. Çalışmada ele alınan sektöre özel veri madenciliği yaklaşımında bir çalışmaya rastlanmaması ile literatürde ki boşluğun doldurulması da amaçlanmaktadır. Ele alınan çalışmada veri madenciliği yöntemlerinden olan karar ağacı tekniğinin C5.0 ve C&R algoritmaları kullanılmış ve sonuçlar karşılaştırılmıştır. C5.0 algoritmasının çalışma mekanizmasında entropi değerlerini ele alırken, C&R algoritması gini algoritmasına göre çalışmaktadır. Çalışma için 727 müşterinin verileri alınmış, 8 tane bağımsız değişken, 1 tanede bağımlı değişken veri setine dâhil edilmiştir. Bağımlı değişken şirkete kayıt olup olmaması ile ilgiliyken, bağımsız değişkende kayıt olanların ya da sadece danışmanlık alanların bilgileri alınmıştır. Bağımız değişken faktörleri: cinsiyet, yaş kategorisi, bölüm, statü, aranılan şehir, yıur dışı deneyimi, başvurduğu program ve aranılan tarih olarak ayrılmıştır. Çalışmada yer alan ilk teknik C5.0 algoritması sonuçlarına göre kayıt durumunu etkileyen en belirleyici bağımsız değişkenler arasında müşterinin mesleki alanı , statüsü , yurt dışı deneyimi yer alırken; ikinci teknik olan C&R algoritması modellemesine göre de müşterinin mesleki alanı, yaşadığı şehir, aranılan tarih bağımsız değişkenleri kayıt durumunu etkilemede en belirleyici faktörler olarak yer almıştır. Çalışma sonucunda SPSS Modeler 18.0 programında yer alan literatürde çok rastlanmayan analysis modülü ile her iki tekniğinde performans değerlendirmesi yapılmıştır. Her iki teknik training ve testing olmak üzere değerlendirildiğinde C5.0 algoritması için trainig doğruluk oranı %60,33, testing doğruluk oranı %46,67 olarak, aynı şekilde C&R algoritması için training doğruluk oranı %57 iken, testing değerlendirmesinde %50,23 değerinde bir başarı oranı hesaplanmıştır.

DETERMINING THE TARGET AUDIENCE BY DATA MINING METHOD IN LANGUAGE EDUCATIONAL INSTITUTIONS ABROAD

SUMMARY

Today, with the development of technology and the innovations it brings, the rate of foreign language learning and its importance has increased significantly. Learning a foreign language in the global order has started to have a critical importance in professional business life, regardless of the sector. This importance has started to be seen more in private sector employees. Due to this importance arising from the private sector, knowing a language has been one of the most determining factors in recruitment. Although university students mostly prefer student exchange programs in terms of both cost and time, this situation has started to be insufficient. For this criterion of great importance, many overseas education consultancy companies have been established in Turkey, which can be used by individuals from all walks of life. These organizations implement programs that appeal to all ages, both educational and working, and traveling. The increase in the number of these organizations day by day has started to cause a competition among the organizations. In this study, it is aimed to learn which criteria affect the registration status by collecting the customer information of the consultancy company discussed in this study. According to this result, defining the customers, determining the target customer group, determining the customer profile that can be offered special offers, changing the customer portfolio when necessary, determining the times when they receive the most customers, it is aimed to present the necessary improvement suggestions. It is aimed to fill the gap in the literature by not finding a study in the sector-specific data mining approach discussed in the study.

The data required for the study were brought together by bringing together the data that the education consultants kept when the customers came for information or when they called for information. The reason for considering the company, which is still new in the sector, aims to progress rapidly due to the fact that it is a new consultancy company in the sector compared to rival companies. Data of 727 customers were taken for the study, 8 independent variables and 1 dependent variable were included in the data set. While the dependent variable was about whether to register with the company or not, the independent variable was the information of those who registered or only received consultancy. Our independent variable factors are divided into: gender, age category, department, status, city sought, overseas experience, program applied for, and date sought.

Gender variable was divided into two groups as male and female, age category has five groups within itself, the field variable is divided into eight groups, the status variable is divided into two groups, depending on whether the individual is a student or an employee. the city variable is divided into five groups within itself, into two groups according to whether the individual has experience abroad or not, the applied program is divided into five groups within itself and the last sought date is divided into 12 groups within itself.

SPSS Modeler program was used in the study and C5.0 and C&R algorithms from decision tree methods were used. The C5.0 algorithm determines entropy values while distinguishing trees and distinguishes according to entropy values. The first method

was made according to the variable area of the first distinction from the C5.0 algorithm model, followed by the searched date, gender and foreign experience criteria. In this case, according to the model output results, the most important factors affecting enrollment are; field, date sought and experience abroad. In particular, the common point of the cluster, which left 70% and 64% of the classification groups as registered, was the fact that the field of the individuals was outside of engineering, being a male individual, and not having experience abroad.

When the results of the two algorithms are compared, it is noteworthy that the determining factors in both algorithms are the field, the city and the searched date. According to the analysis result of the C5.0 algorithm, when the performance value of the model was interpreted for the test data, it correctly interpreted 55 unregistered persons as unenrolled, and misinterpreted 31 unregistered persons as enrolled. It correctly classified 29 unregistered persons as not registered as not registered, while it correctly classified 65 unregistered persons as registered. In this case, while the model achieved 69.74% success for the training part, it achieved a 46.67% success rate for the test data. According to the C&R algorithm analysis result, when the performance value of the model is interpreted for the test data, it correctly interprets 51 people who did not register as not registered, misinterpreted 57 people who did not register as registered. It correctly classified 56 unregistered people as not registered as not registered, and correctly classified 52 unregistered people as registered. In this case, 57.22% success was achieved for the model training part, while a 50.23 percent success was achieved for the test data.

1. GİRİŞ

Dünya her geçen gün yaşanan dijitalleşme süreçleri ile birlikte sürekli bir değişim ve gelişim içindedir. Bu değişimin getirdiği yenilik ile iş dünyasında fiziksel sınır kalmamıştır. Dünyanın bir ucunda gerçekleştirilen bir faaliyet ya da hizmet, diğer ucunda anında duyulmaya başlanmış ve iş uygulamalarını kolaylıkla yürütülebilir hale getirmiştir. Özellikle 2019 yılının Aralık ayından itibaren dünyada kendini hissettiren ve sonrasında hayatın her alanında etkili olan pandemi tüm iş yapma biçimlerini değiştirmiştir. Eğitim sektörü de özellikle pandemi ile birlikte ilk uzaktan modele geçiş yapmak zorunda kalınan sektörlerden biri haline gelmiştir. Yükseköğretim kurumlarında 12 Mart 2020'de diğer okullar gibi eğitim öğretime ara verilmiş, 23 Mart 2020 tarihinden itibaren 2019-2020 öğretim yılı bahar dönemi eğitim öğretim dijital ortamda, açık ve uzaktan öğrenme sistemi ile sürdürülmesine karar verilmiştir [1].

Dil eğitimi de eğitim sektörünün getirmiş olduğu online sistem altında son yıllarda her geçen gün önemi artan bir konu olmaya başlamıştır. Özellikle uluslararası bağlantılı iş dünyasında aktif rol almak isteyenler için ekstra bir ilgi görmektedir. Aynı zamanda işletmelerin artık dil bilen çalışan araması da bu durumu bir bakımdan zorunlu hale getirmiştir. Bu durumun zorunlu bir hale gelmesi de işe alımlarda en önemli kriterler arasında olmasına sebebiyet vermesine başlamıştır. İyi derecede dil bilgisi olan bir çalışanın özellikle özel sektörde terfi alması benzer kariyer geçmişine sahip başka bir çalışana oranla daha yüksek olduğu gözlemlenmiştir.

Ülkemizde yabancı dil eğitimi şu anda hem 4+4+4 sistemi kapsamında ilkokul, ortaokul ve liselerde hem de yükseköğretim düzeyinde üniversitelerde verilmektedir. Yabancı dil olarak ağırlıklı İngilizcenin öğretildiği ülkemizde, özellikle lise ve üniversitelerde Almanca, Fransızca, Rusça, Arapça, Çince, İtalyanca, İspanyolca gibi diller de isteğe bağlı olarak öğretilmektedir. Hatta yükseköğretim düzeyinde, ODTÜ ve Boğaziçi Üniversitesi gibi üniversiteler yıllardır eğitim dili tamamen İngilizce olarak faaliyetlerini sürdürmektedir. Eğitim dili Türkçe olan üniversitelerde de bazı bölümler yüzde yüz İngilizce (örneğin; İngilizce Tıp) bazı bölümler de ise yüzde otuz İngilizce eğitim vermektedir. Eğitimi tamamen Türkçe olan bölümlerde ise genellikle

1. sınıfta haftada en az 2 saat olmak üzere genel İngilizce eğitimi verilir. Meslek yüksekokullarındakiler dâhil üniversitelerdeki bölümlerin çoğunda ilerleyen dönemlerde Mesleki İngilizce dersi de verilmektedir. Yukarıda bahsedilen kronolojik gelişmeler ışığında Türkiye'de genelde yabancı dil eğitiminin özelde ise İngilizce öğretiminin nereden nereye geldiğine dair bir çerçeve ortaya koymaya çalıştık. Görünen o ki öğrenciler eğitim hayatlarının neredeyse tamamında İngilizceye temas etmektedir. Mevcut yetişkinlerin önemli bir kısmı da kendi imkânlarıyla özel dersler, kursalar, bireysel çalışmalar gibi yollarla İngilizce öğrenmeye çabalamaktadır [2].

Nitekim 2020 yılı için uluslararası İngilizce yeterlik indeksi incelendiğinde; 100 ülke içinde birinci sırada olan Hollanda'nın çok yüksek (very high proficiency) İngilizce seviyesine sahip olduğu, Türkiye'nin ise 69. sırada düşük (low proficiency) İngilizce seviyesine sahip ülkeler içinde olduğu görülmektedir [3].

Sezer'e göre [4] Türkiye'de İngilizce öğretimine önem verilmesinin nedeni Türkiye'nin ekonomik ve teknolojik alanda gelişmek istemesiyle açıklanabilir; kültürel, ekonomik ve teknolojik olarak diğer ülkelerle ilişkileri yoğunlaşan Türkiye, bu ilişkileri İngilizce ile sürdürmek durumundadır. Bu zorunluluk yüzünden, İngilizce öğretimi uzun zamandır Türkiye'de bir seferberlik halini almıştır.

Bu derece önemli olan bir konuda ülkemizde ilköğretimden itibaren verilen eğitim yeterli olmamaktadır. Bu yetersizliği önlemek amacıyla bir takım geliştirmelerde bulunulmuştur. Bunlardan bir tanesi Ortadoğu Teknik Üniversitesi, Boğaziçi Üniversitesi ve diğer [bazı] üniversitelerde [öğretimin tamamı], bazı üniversitelerin kimi bölümlerinde ise öğretimin üçte biri İngilizce yapılmaya başlanmıştır [5].

Dil gelişimi amacıyla da açılan yurt dışı eğitim ve danışmanlık kurumları özellikle son 10 yıldır çok fazla sayıda yurt dışında dil okulu, lisans, yüksek lisans eğitimleri, yaz kampı ve buna benzer programlar düzenlemektedirler. Ayrıca online sisteme geçiş ile birlikte özellikle yurt dışı dil okulu paketlerinin uzaktan yürütülebilir hale getirilmesi de bu durumun avantajlarından biri haline gelmiştir. Gün geçtikçe de dil öğrenimine talebin artması yurt dışı eğitim danışmanlık şirketleri arasında da bir rekabete yol açmıştır. Eğitim şirketlerine artan rağbet ve buna bağlı oluşan rekabet üzerine kurumlar iyileştirme çalışmalarına gitmiş, gerek müşteri portföyünü gerek kurumun içerdiği eğitim paketlerini geniş tutmaya başlamıştır. Fakat günümüzde bu kurumlara

başvuruların sadece eğitim programları ile sınırlı kalmayıp vize başvurusunda bulunanlarında da sayısı arttığından vize aracılık hizmetleri artmıştır. Artan bu sayı üzerine de sadece vize danışmanlığı veren yurt dışı danışmanlarının sayısını da arttırmıştır. Ayrıca yurt dışı eğitim danışmanı olmak isteyen kişiler için de yurt dışı eğitim kurumları en az bir kere yurt dışında eğitim alma şartı koymuşlardır. Yurt dışı deneyimin sadece dil öğrenimi dışında faklı avantajları da yer almaktadır.

Öğrencilerin yurt dışında eğitimi arzulamasında kültürel anlamda yurt dışı deneyim kazanma, yeni bir yaşam biçimi deneyimleme ve girilen yeni çevrede karşılaştıkları sorunlarla başa çıkarak kazanımlar edinme etkili olan hususlardır [6]. Uluslararası öğrencilerin yurt dışı eğitim deneyimini hedef ülke tercihi kadar hedef yükseköğretim kurumu da etkiler. Mazzarol ve Soutar uluslararası öğrencilerin eğitim görmek istedikleri hedef ülkedeki üniversite tercihinde üniversitenin itibarı, önceden elde edilmiş dereceleri kabul etmesi, niteliğinin yüksek olması ve uzman öğretim üyelerinin bulunması, aşina olunan üniversitelerle bağlantılarının olması, uluslararası öğrenci sayısının fazla olması ve üniversite hakkında bilgi edinilebilecek mezunlarının olması etkenlerinin üniversite tercihini etkilediğini tartışmıştır [7]. Gerek öğrencilerin gerek çalışanların özellikle üniversite tercihlerinde dikkat etmesi gereken kriterlerin fazla olması yurt dışı eğitim kurumlarına her anlamda büyük bir iş düştüğünü göstermektedir.

Bu çalışmada da amaç ele alınan yurt dışı danışmalık şirketinde hizmet alan müşteri gruplarının belirlenmesi ve yurt dışı eğitim kurumları ile ilgili konularda çok fazla çalışmaya rastlanmadığı için literatürün zenginleştirilmesi amaçlanmıştır.

2. LİTERATÜR ARAŞTIRMASI

Bu bölümde veri madenciliği metotlarından karar ağacı tekniği, farklı sektörlerden müşteri profili belirleme çalışmaları ve yabancı dilin önemine dair literatürde yer alan çalışmalara yer verilmiştir.

Karabulut [8] yaptığı çalışmada özel bir sağlık kurumundan hizmet alan hastaların verilerini kullanarak veri madenciliği tekniği ile bir veri analizi ve çıkan sonuçlara göre iyileştirme önerisi sunmayı amaçlamıştır. Kurumdan hizmet alan bireylerin özelliklerini ve hizmet alınan şubelerin nitelikleri göz önünde bulundurularak bir hasta profili çıkarılması hedeflenmiştir. Çalışmada veri madenciliği metotlarından karar ağacı ve algoritmaları kullanılmıştır. Veri setinde 10 farklı değişken sunulmuş, 340.900 hasta kaydı verisi kullanılmış yöntem olarak da 3 farklı algoritma ile modeller kurulmuş ve çıkan sonuçlar karşılaştırılmıştır. Veri madencilği yazılımlarından IBM SPSS Modeler kullanılmış ve Chaid, Quest ve C5.0 algoritmaları kullanılmıştır.

Can ve Gerşil [9] birlikte yaptıkları bir çalışmada günümüzde teknolojinin hızlı bir şekilde gelişmesi, insanların yoğun bir tempoda çalışması, zamanın kısıtlı olması ve bunun benzeri durumlardan ötürü müşterilerin büyük bir çoğunluğunun online alışverişi tercih etmesi göz önünde bulundurularak e-ticaret sektöründe bir çalışma yapmışlardır. Çalışmada 12330 adet veri alınmış, incelenen verilere veri madenciliği sınıflandırma tekniklerinden karar ağacı algoritmalarından Chaid, Quest, C&RT ve C5.0 algoritmalarının hepsi denenmiştir. Gerçekleştirilen analizlerde online alışveriş davranışlarında en önemli etkenlerin ziyaret edilen ay ve sitenin özel bir gün ziyaret zamanına yakınlığı olduğu sonucuna varmışlardır. Hafta sonları, en çok satış yapılan aylarda ve özel günlerde müşterilere özel kampanyalar yapılmasının satışları olumlu yönde etkileyeceği yönünde çözüm önerisi sunulmuştur.

Yıldıztepe ve Kocataş [10] son yıllarda ülkemizde hatta dünya genelinde giderek artan işsizlik sorunu üzerine sağlıklı bir politika oluşturabilmek için bu konu üzerinde bir veri derlemesi yapıp veri madenciliği çalışması yapmıştır. Bu çalışmada Türkiye'nin iş gücü hakkında bir değerlendirmede bulunulması amaçlanmıştır. 2013 yılında iş hayatında bulunan fakat bir yıl önce (2012'de) çalışmadığını belirten kişilerin,

2013'deki iş durumu hedef değişken olarak belirlenerek CART ve C5.0 algoritmaları kullanılarak bir karar ağacı modeli oluşturulmuştur. İş durumunu bu şekilde ifade eden eden kişilerden rastgele 22.206 kişi seçilmiştir. Modeli oluşturabilmek için IBM SPSS Modeler programı kullanılmıştır. İki model sonucuna göre cinsiyet ve mezuniyet durumu iş gücü durumunu etkileyen ortak faktörlerdir.

Aytekin ve ark. [11] bir işletmenin veri tabanından müşteri yorumlarından örneklem seçerek karar ağacı metodu kullanarak müşteri yorumlarını şikâyet-talep-teşekkür sınıflarına ayıracak bir metin sınıflandırma uygulaması yapmışlardır. Online alışveriş yapılan bir siteden 22 müşteri yorumu incelenerek ele alınan yorumlardan belirledikleri sınıfları temsil edebilecek nitelikteki kelimeler çıkarılmış. Araştırma örnekleminde şikâyet-talep-teşekkür sınıflarını temsil edebilecek kelimeler sınıflara atanmıştır ve tüm ihtimaller hesaba katılarak bir karar ağacı çıkarılmıştır. Çalışmada modele ilişkin bir kod da yazılabileceği belirtilmiştir.

Emel & Taşkın [12] veri madenciliği yöntemlerinden C&R karar ağacı tekniğini kullanarak bir perakende işletmesinin satış analizini gerçekleştirmişlerdir. Bilgilere göre, perakendeci doğrudan satış dışında telefon ve internet üzerinden satış gerçekleştirmektedir. Çok kanallı pazarlama yöntemi kullanan işletmenin öncelik vermesi gereken hedef müşteri gruplarını belirlemeyi amaçlamışlardır. Analiz için C&R karar ağacı tekniği, yazılım programı için Clementine v8.1 yazılımından yararlanılmış. Veri tabanından 8000 adet müşterinin kaydı çıkarılmış. Müşteri numarası, müşteri durumu, bulunduğu şehir gibi değişken faktörler belirlenmiştir. Çalışmanın sonunda ürün satışı ve müşteri satın alma durumu arasında bir analiz kurularak ve müşteri profili çıkarılarak pazarlama stratejisi geliştirilmesi amaçlanmıştır.

Çakır & Kamal [13] çalışmalarında Ulaştırma ve Altyapı Bakanlığı Ana Arama Kurtarma Koordinasyon Merkezi veri tabanından 2001-2016 yılları arasında meydana gelen ticari yük gemi kazalarının kayıtlarına ulaşarak 1091 deniz kazasından gerekli değişkenler göz önünde bulundurulup, filtrelenerek sadece 535 ticari geminin analizi yapılmıştır. Belirli bölgelerde meydana gemi kazaları CHAİD karar ağacı yöntemi kullanılarak incelenmiştir. CHAİD karar ağacı büyük miktarda veri yığınlarını analiz etmede kullanılan en yaygın yöntemlerden biri olduğu için bu teknik seçilmiştir. Kazaların türü, gemi faktörleri, zaman faktörleri ve diğer faktörler arasındaki ilişki inceleme alınmıştır. Çalışmada kullanılan değişkenler gemi tipi, gemi yaşı, gemi

büyüklük durumu, kaza zamanı, kaza mevsimi, kaza nedeni gibi değişkenlere ayrılmış ve bu değişkenlerde alt değerlere ayrılmıştır. CHAID Karar Ağacı yöntemi ile ticari yük gemileri için çıktı değişkeni olan kaza tipi ile gemi tipi, kaza zamanı, kaza sektörü vs. gibi girdi değişkenleri arasındaki ilişkiler incelenmiştir ve en önemli girdi değişkenleri olarak kazanın olduğu bölge gemiye pilot alınması durumu, gemi tipi ve kaza zamanı tespit edilmiştir.

Kadirhanoğulları ve ark. [14] birlikte yaptıkları çalışmada Iğdır ili Merkez ilçesinde kentsel bir bölgesinde, örnekleme yöntemiyle seçilen kişilerin organik gıda ürünlerinin tüketimi hakkında bilgi düzeylerini, tercihlerini ve düşüncelerinin belirlenmesi amacıyla; yüzyüze görüşme ile yapılan anket verileri derlenerek bir çalışma yapılmıştır. Organik gıda ve tarım ürünlerini satın alma davranışlarının belirlenmesi amaçlanmaktadır. Çalışma boyunca toplam 168 kişi ile anket yapılmış, çalışmada veri madenciliği kullanılmış ve modelleme boyunca karar ağacı tekniği uygulanmıştır. Yapılan modellemede bağımlı değişken olarak satın alma, bağımsız değişkenler olarak eğitim durumu, yaş, cinsiyet, medeni hal, meslek durumu vb bağımsız değişkenler kullanılmıştır. SPSS'te CHAID algoritması ile karar ağacı oluşturulup sonuçlar yorumlanmıştır. Çıkan sonuçlara göre hanedeki birey sayısı, hane halkı gelir durumu, cinsiyet gibi faktörlerin organik gıda tüketiminden önemli rol oynadığı belirtilmiştir.

Sevüktekin ve ark. [15] veri madenciliğinde farklı bir konuyu ele almışlardır. Bursa Emniyet Müdürlüğü veri tabanından suçlulara ait bilgiler alınmıştır. İlgili veri tabanından suçlulara ait bir profil çıkarılmaya çalışılmıştır. Çalışma sonunda ağır suç işlenmesinde en çok etkili faktörlerin tespit edilmesi amaçlanmaktadır. Emniyet veri tabanından 34.521 kişinin verilerin alınmıştır ve kullanılan değişkenler şuçun işlendiği saat, suçun işlendiği bölge, suçu işleyen bireyin cinsiyeti, bireyin medeni hali gibi toplam 9 faktör belirlenmiştir. Analizde karar ağacı algoritmalarından CHAID algoritması kullanılarak kişilerin ağır suç işlemesinde en etkili bağımsız faktörlerin olay saati, doğum yeri, meslek değişkenleri olduğu görülmüştür.

Bardi & Can [16] yaptıkları çalışmada ülkemizde KOBİ kapsamında faaliyet gösteren işletmeleri ele almışlardır. Çalışmada ki amaç finansal başarısızlığa neden olan finansal oranları belirlemektir. Diskriminant analizi ve C5.0 karar ağacı algoritması kullanılmıştır. İşletme burada finansal başarısızlık kapsamında üst üste 2 sene başarısız olan işletmeleri ele almıştır. 20 finansal olarak başarılı, 20 finansal olarak başarısız işletme seçilmiştir ve belirli yıllar arasındaki finansal bilgileri kullanılmıştır.

Diskriminant analizine göre firmaların başarılı ve başarısız olarak gruplara ayrılmasında en etkili faktörlerden biri kaldıraç oranıdır. Çünkü çalışmaya göre çıkan sonuçlarda kaldıraç oranı ve firma karlılığı arasında ters orantı olduğu fark edilmiştir. C5.0 algoritmasına göre karar ağacı oluşturulurken en güçlü etkiye sahip olan faktör brüt kar marjı ve kaldıraç oranı faktörleridir. Brüt kâr marjı 0,12'den büyük ve kaldıraç oranı 0,26'ya eşit veya küçük olan 29 firmanın %100'u finansal başarılıdır. C5.0 karar ağacı buna benzer kurallar oluşturmuştur. Her iki modelin ortak sonuçlarına göre kaldıraç oranı, brüt kar marjı, firmaların yaşı gibi faktörler firmaların gruplara bölünmesinde etkili bağımsız faktörler olduğu belirlenmiştir.

Arslantürk Çöllü ve ark. [17] yaptıkları çalışmada 2016 ve 2018 yılları arasındaki Borsa İstanbul' kayıtlı işletmeleri ele alarak bu işletmelerin finansal başarısızlarındaki etkili faktörleri tespit etmektir. Bunun için veri madenciliği tekniklerinden karar ağacı ve algoritmalarını kullanmışlardır. Öncelikli şirketlerin finansal durumunu başarılı ve başarısız olarak değerlendirebilmek için Altman Z skoru kullanılmış. Hesaplanan Z skoru neticesinde 2,99 dan büyük çıktığı takdirde finansal olarak güvenli bölgede sayılıp başarılı kategorisine alınmıştır. 1,81 ve 2,99 arasında olanların gri bölgede olması 2,99 dan küçük olanları tehlikede bölgede olmasına ve bu iki bölgenin başarısız kategorisinde yer almasına karar verilmiştir. Kategorilere ayrıldıktan sonra finansal başarısızlıkları finansal oranlar yardımıyla (Cari oran, Nakit oranı vb) ve oranlar ve kategoriler arasındaki en uygun sınıflandırma algoritmasını bulmak için CHAID, Exh-CHAID, CART ve QUEST algoritmaları kullanılmıştır. Uygulama sonunda kullanılan algoritmalar arasından en doğru sınıflandırmayı % 95 doğruluk oranında CART algoritmasının yaptığı sonucuna varılmıştır.

Koçak [18] yaptığı bir çalışmada örgütsel bağlılığı ele almıştır. Çalışanların kurumlarına bağlı olmasının, kurumun ilerlemesi için daha özverili bir performans sergileyeceği ve kurumda çalışmaya devam etmek için daha istekli olabileceği düşünülerek örgütsel bağlılığın önemi üzerine bir çalışma yapmıştır ve veri madenciliği metotlarını kullanarak örgütsel bağlılığı tahmin etmeyi amaçlamışır. Kamu ve özel sektörde çalışan toplam 526 kişiye uygulanan, toplam 4 bölüm içeren ve 39 adet sorudan oluşan anket uygulanmıştır. Bu sorulara 5'li likert ölçeği kullanılmıştır. Çalışmada veri madenciliği metotlarından karar ağacını algoritmasını uygulamak için Python programlama dili uygulanmıştır. Bu programlama dilinde karar ağacı algoritmalarında CART algoritmasının kullanmıştır ve veri test-eğitim

aşamasında Python programlama diline ait Scikit-Learn makine öğrenme algoritması kullanılmıştır. Öncelikle ilk karar ağacı uygulamasından sonra algoritmaya overfitting (budama) işlemi uygulanmıştır. Bu işlem oluşturulduktan sonra uygulanan karar ağacında daha az kuralın ortaya çıktığı ve bağımsız değişkenlerin hepsinin karar ağacına etki etmediği fark edilmiştir.

Yakut & Korkmaz [19] 'İnsani Gelişmişlik Endeksinin Karar Ağacı Algoritmaları ile Modellenmesi: BM'de Bir Uygulama 2010-2017 Dönemi' isimli çalışmasında İGE değerini göz önünde bulundurarak bir çalışma yapmıştır. İGE insan hayatını ilgilendiren bir değerlendirme ve hesaplama oranıdır. İnsan faktörü söz konusu olduğunda önemli olan bu değerin oranına göre ülkeler; gelişmiş, az gelişmiş, orta düzey gibi kategorilere ayrılmaktadır. Bu çalışmada da aaç İGE ye etki eden faktörlerin önem derecesine bakmak ve belirli bir kural tabanı oluşturmaktır. Çalışmada veri madenciliği tekniği karar ağacından C5.0 ve Gini algoritması kullanılmıştır. Birleşmiş Milletler Kalkınma Programından 2010 ve 2017 yılları arasındaki 79 ülkenin verilerine ulaşılmış ve bu ülkeler İGE'ye göre sınıflandırılmıştır. Sınıflandırılan İGE değerlerine göre İGE'ye etki eden faktörleri belirlemek ve karar ağacını oluşturabilmek için 15 bağımsız değişken belirlenmiştir. Belirlenen değişkenleri veri ön işlemeden geçirme işleminde spearman korelasyon analizi işlemi gerçekleştirilmiş ve belli bir değerin üstünde yüksek korelasyon değerine sahip bağımsız değişkenler karar ağacı analizinden çıkarılmıştır. Verilerin %75'i eğitilmiş, geriye kalan kısım ise test için kullanılmıştır. Sonuçlara göre Gini Algoritması ile oluşturulan karar ağacına göre sınıflandırma en önemli faktör eğitim endeksi, C5. Algoritmasıyla oluşturulan karar ağacına göre de en önemli faktör eğitim endeksi olarak belirlenmiştir.

Beşli & Tenekeci [20] yeryüzündeki tüm varlıklar için önemli olan ormanlarımızı yangınlardan korumak amacıyla bir orman yangını tahmini çalışması yapmışlardır. Ne kadar erken tahmin edilirse önlem almanın o kadar kolay olduğunu düşünmüşlerdir. Orman yangını için kullanılmış veriler için Kanada Bölgesinde 2013 ve 2014 yıllarında ki gerçekleşen yangınların verileri kullanılmıştır. Bu tarihler arasındaki verilerin eğitim, test işlemleri gerçekleştirilmiştir. Verilen bilgilere göre veriler Terra ve Aqua uydularında bulunan MODIS (Moderate Resolution Imaging Spectroradiometer) algılayıcıdan alınmıştır. MODIS tarafından veriler kaydedilmiştir ve Nasa bu verileri ulaşıma açmıştır. Çalışmada bu tahmini yapmak için kullanılan

parametreler NDVI indeksi, LST değeri ve TA parametreleridir. NDVI parametresi bitki örtüsü hakkında bilgi taşır, LST parametresi hedef bölgedeki toprağın sıcaklığı ile ilgili bilgi verir. TA parametresi ise direkt yangın ile ilgili bilgi verir ve diğer parametrelere göre daha güçlü bilgiler verir. Veri toplama ve ön işleme işleminden sonra 804 verinin % 70 i eğitim örneği için ayrılmış geriye kalan ise test için ayrılmıştır. Karar ağacı sonuçlarına göre duyarlılık oranı % 98.62 çıkmıştır. Karşılaştırma yapma amacıyla başka metotlar ile de tahmin yapılıp duyarlılık

Okatan & Işık [21] 'Sağlık Harcamalarının Tahmininde Karar Ağacının Kullanımı' isimli çalışmasında ülkemizde sağlık alanına büyük miktarda kaynak ayrıldığını düşünerek bu ayrılan kaynağın bütçe planlamasını ve analizinin sağlıklı bir şekilde yapılması gerektiğini düşünmektedirler. Sağlık harcamaları aynı zamanda sigorta şirketleri içinde önemli bir husustur çünkü bu durum detaylı bir şekilde analiz edilirse sigorta şirketlerinin suiistimal edilmesinden de korunmuş olur. Bu çalışmada sağlık harcamaları tahmini için veri madenciliği metotlarından karar ağacı kullanılmıştır. İlk aşamada veri ön izleme için normalizayon uygulanmıştır ve bunu için aralık ölçeklendirme (min-max normalization) kullanılmıştır. Karar ağacı için kullanılan veriler sigara kullanımı, bölge, yaş, Vücut Kitle İndeksi, Çocuk Sayısı, Sağlık Harcamasıdır. Toplam 1400 kişinin verisi kullanılmıştır. Verilerin karar ağacına göre uyarlanabilmesi için sınıfsal veriler sayısal verilere dönüştürülmüştür. Çalışmada RapidMiner isimli veri madenciliği yazılımı kullanılmıştır. Algoritma olarak C4.5 algoritması tercih edilmiştir. Veri kümesinde normalizasyon işlemi yapılırken 3 farklı yöntem denenmiştir. Önce veriler normalize edilmemiştir, daha sonra 0-1 arasında normalize edilmiş ve son olarak 0-100 arasında normalize edilmiştir. NRMSE değeri normalleştirilmemiş veri kümesinde en düşük değere sahip olduğu için daha iyi tahmin yapıldığı söylenebilir. Sonuç olarak öğrenme verilerin % 70 i öğrenme, % 30 u ise test için ayrılmıştır. Çalışma sonunda farklı normalizasyonlar ile yapılan denemelerde normalizasyon yapılmaması durumunda daha iyi sonuçların çıktığı görülmüştür.

Demirel&Yakut [22] 'Karar Ağacı Algoritmaları ve Çocuk İşçiliği Üzerine Bir Uygulama' isimli çalışmasında çocuk işçiliğine sebebiyet veren parametreleri veri madenciliği yönteminden karar ağacı algoritmaları ile belirlemektir. Bu durumun dünya üzerinde önemli bir konu olduğu düşünülmektedir. Çalışmada gelişmekte olan ülkelerde 114 tanesi seçilmiştir. Diğer ülkelere ait veriler eksik olduğu için çalışmadan çıkarılmıştır. Karar ağacı algoritmalarından CART ve CHAID kullanılmıştır.

Kullanılan faktörler ise; SGP, GSYH, Yoksulluk oranı, İşsizlik oranı, Kentsel oran, Nüfus, Yetişkin okur-yazar oranı, Göç Oranı, Gini Endeksi'dir. Bu parametreler literatürden yararlanılarak da belirlenmiştir. Öncelikli olarak regresyon ağacında hangi ağaç yapısının daha iyi olduğunu anlayabilmek için her iki algoritmaya da 3 farklı yöntem uygulanmıştır. İlk olarak verilerin % 70 i eğitim %30 u test daha sonra % 50 eğitim % 50 test ve son olarak %30 eğitim % 70 test olarak ayrılmıştır. İlk yapılan denemenin daha iyi sonuç verdiği görülmüştür. CART Algoritmasına göre bulunan bulgularda çocuk işçiliğini en çok etkileyen faktör SGP faktörü, CHAID algoritmasına göre de en etkileyeci faktör SGP olarak belirlenmiştir. İki algoritmaya göre de en çok etkileyen faktör aynı çıksa da farklılıklar bulunmuştur. CART algoritması düğümleri iki alt dala bölerek yapsa da CHAID algoritması ikiden fazla gruplara bölmüştür. CHAID algoritması ile sonuçların daha kapsamlı görülebileceği sonucuna varılmıştır.

Akbal ve ark. [23] 'Karar Ağaçları ile Telefon Dolandırıcılığı Verilerinin Analizi' isimli çalışmasında telefon dolandırıcılığı verilerini analiz etmeyi amaçlamıştır. Bunu da veri madenciliği yöntemlerinden karar ağacı metodundan C4.5 algoritmasını kullanmıştır. Çalışma için toplam 115 kişinin verisi alınmış, 75 i eğitim 40 ı ise test verisi olarak ayrılmıştır. Kullanıcılara hem bireysel hem de telefon dolandırıcılığı üzerine sorular sorulmuştur. Bireysel sorular için öğrenim durumu, yaşam yeri, meslek, yaş, aylık gelir parametreleri belirlenmiş bu parametrelerde alt dallara ayrılmıştır. Telefon dolandırıcılığı ile ilgili sorularda evet/hayır şeklinde cevap verilebilecek sorular yönlendirmişlerdir. Çalışma sonunda insanların bireysel özelliklerine bakılmaksızın kişilerin telefon dolandırıcılığı konusunda yeterli bilgiye sahip olmadıkları sonucuna varılmıştır. Çıkan sonuçlara göre insanlar böyle bir durumla maruz kaldığında gerekli prosedürleri bilmediği, bu durumla karşı karşıya kalan insanların şikayetde bulundukları zaman bir çözüm bulamadıkları sonucuna varılmıştır. Aksu & Karaman [24] 'Karar Ağaçları ile Bir Web Sitesinde Link Analizi ve Tespiti' adlı çalışmalarında bir web sitenin ana sayfasında olması gereken linklerin veri madenciliği ile analizini gerçekleştirmeyi amaçlamışlardır. Bu şekilde ana sayfanın daha kullanışlı ve siteye girenler için daha kolay olacağını düşünmüşlerdir. Çalışmada kullanılan veriler bir üniversitenin web sitesinin 2015-2017 yılları arasındaki verileri kullanılmıştır. Bu parametreler arasında girilen linklerin; sayfa görüntüleme sayısı, sayfada kalma süresi vb. parametreler bulunmaktadır. Duyuru linkleri, etkinlik linkleri vb. linkler çalışmaya katılmamış, toplam 2536 veri

kullanılmıştır ve veriler veri madenciliği yapılacak şekilde düzenlenmiştir. Veriler Weka programı yardımıyla C4.5, Random Forest, REP Tree, Logistic Model Tree ve CART algoritmaları ile sınıflandırılmıştır. Doğru sınıflandırma oranı en yüksek olan C4.5 algoritması ile karar ağacı oluşturulmuştur. Oluşturulan karar ağacı ile web sitesinde bulunması gereken linkler (öge) belirlenmiştir. Ersöz ve ark. [25] 'Tüketicilerin Cep Telefonu Tercihlerinin Karar Ağacı İle Modellenmesi' isimli çalışmalarında müşteriler cep telefonu alırken hangi parametrelere dikkat ettiğine yönelik bir çalışma yapmışlardır. Bu araştırma için karar ağacı metodunu kullanıp CART algoritmasını tercih etmişlerdir. İlk olarak bir e-ticaret sitesinin en çok satan telefon modellerine ait verilere erişilmiş ve bir anket hazırlanmıştır. Bu ankette cep telefonlarına ait özelliklere yer verilmiş ve toplam kişiye anket uygulanmıştır. Ankette 'tercih ederim', 'etmem' ve 'kararsızım' ifadeleri yer almıştır ve rastgele seçilen kişilere anket uygulanmıştır. Bir telefon modeli için hangi ifade en çok tercih edilmişse o ifade o telefonun etiketi olmuştur. Veriler hazırlandıktan sonra CART algoritması ile ağaç oluşturulmuştur. Çalışma sonunda alıcılar cep telefonu seçerken en önemli faktörün fiyat olduğunu belirlemişlerdir.

Pala ve ark. [26] 'Meme Kanserinin Teşhis Edilmesinde Karar Ağacı ve KNN Algoritmalarının Karşılaştırmalı Başarım Analizi' isimli çalışmalarında makine öğrenmesini medikal alanda uygulamışlardır. Meme kanseri günümüzde ölüm oranının en yüksek olduğu kanser türlerinden biri olduğu için erken teşhisin önemli olduğunu vurgulamışlardır. Çalışmalarında En Yakın Komşu Algoritması ve karar ağacı algoritması yazılım olarak Python kullanılmıştır. Çalışma sonunda En Yakın Komşu Algoritmasının % 97 doğruluk oranı ile daha iyi sonuçlar verdiğini gözlemlemişlerdir.

Çalış ve ark. [27] 'Veri Madenciliğinde Karar Ağacı Algoritmaları İle Bilgisayar Ve İnternet Güvenliği Üzerine Bir Uygulama' çalışmalarında bilgisayar ve internet güvenliği üzerine veri madenciliği uygulaması yapmışlardır. Çalışmada 10 soruluk bir anket uygulanmış. Sorularda uygulanan kişilere ait yaş, cinsiyet, eğitim durumu ve internet kullanımı ile ilgili bilgiler yer almıştır. Anket farklı özelliklere sahip 300 kişiye uygulanmıştır. Veri toplama işleminden sonra SPSS Clementine programında C5.0, C&RT, CHAID ve QUEST algoritmaları uygulanmış ve % 81,67 doğruluk oranı ile C5.0 algoritması seçilmiştir ve bu algoritma ile karar ağacı oluşturulmuştur.

Çalış [28] ‘Veri Madenciliği Yaklaşımı İle Bireysel Müşterilerin Kredi Ödeme Performanslarının Değerlendirilmesi’ araştırmasında Türkiye’nin önde gelen bir bankasının bir şubesine ait kişisel ve kredi bilgileri alınarak müşterilerin kredi geri ödeme performansı ile ilgili bir çalışma yapmıştır. Bankaların kredi verme durumu bankalar için önem ve risk arz eden bir durum olduğu düşünülerek bu konu hakkında karar verme durumunu en az riskle almak için bir veri madenciliği uygulaması yapmayı, kredi geri öderken hangi değişkenlerin önemli olduğunu bulmayı amaçlamışlardır. Banka veri sisteminden müşterilere ait kişisel bilgilere ulaşıp (yaş, cinsiyet, medeni durum vb.) toplam 12 adet parametre belirlemişlerdir ve 200 müşterinin bilgisini alarak analiz gerçekleştirmişlerdir. Kredi ödemesini belirli aralıklarla aksatan müşteriler kanuni takip olarak, düzenli ödeyenler ise normal ödeme olarak ayırmışlardır. Çalışmaya önce k-means algoritması uygulanmış ve veriler kümelenmiştir. Müşteri profilini belirlemek için de SPSS Clementine programı kullanılmış ve C&RT, C5.0, CHAID ve QUEST algoritmaları uygulanmıştır. Bu uygulama sonunda müşterilerin kanuni takibe düşmesini azaltmayı amaçlamışlardır.

Altunkaya [29] ülkelerin kredi notu derecelendirmesinde etkili olan faktörleri belirlemeyi amaçlamışlardır. Kredi derecelendirme çalışması için 27 değişken (cari işlemler dengesi, enflasyon, tüfe, gelir vb.) belirlenmiştir. Veriler belirli yıllar arasındaki ülke kredi notlarının 3 farklı kuruluşun derecelendirme sonucuna göre ele alınmıştır, bunun açıklaması üç farklı kuruma göre kredi derecelendirmeye etki eden faktörlerin belirlenmesidir. Analiz için CART, CHAID, QUEST, C5.0 ve Neural Network metotları kullanılmıştır. Örneğin CART algoritmasının 2012 yılı için Fitch kuruluşuna göre önemli üç parametresi enflasyon tüfe, cari işlemler dengesi ve kredi derinliği bilgi endeksi olarak bulunmuştur. Analiz sonucunda 3 farklı kurumun derecelendirmelerinin farklı modeller ile kurulduğunda buna bağlı olarak başarı oranının değiştiği görülmüştür. Örneğin Moody kurumu tarafından belirlenen kredi notlarının CART algoritmasına göre daha yüksek bir başarı oranı gösterdiği sonucuna varılmıştır.

Dolgun & Ersel [30] çalışmalarında bankacılık sektöründe rekabetin arttığını varsayarak doğrudan pazarlama kampanyalarına etki eden parametreleri belirlemeyi amaçlamışlardır. Çalışmada bir ülkede belirli tarihler arasında 45211 kişinin verisi alınarak 16 bağımsız parametre (yaş, meslek, medeni durum… vb) ve 1 bağımlı değişken kullanılmıştır. Analiz için C&R, CHAID, Lojistik regresyon, TAN, MB,

RBF, Linear metotları yazılım programı olarak IBM SPSS kullanılmıştır. Kurulan modellerin başarı performansını ölçmek için duyarlılık, genel başarı, seçicilik, Matthews Korelasyon ve f ölçütü değerleri belirlenmiştir. Sonuç olarak bir kampanyaya göre yola çıkılan bu problem de 7 farklı metot ve 5 farklı değerlendirme ölçütüne göre çıkarım yapmışlardır. F ölçütü, Mathews korelasyon katsayısına göre CHAID algoritması ve lojistik regresyonun başarılı olduğu sonucuna varmışlardır. Bütün metotların ortak sonuçlarına göre ortak 3 parametre belirlemişlerdir.

Dil eğitimi ve profil belirleme ortak çalışmalar;

Erbay ve ark. [31] çalışmalarında dil eğitimi veren kurumda öğretmen profili belirleme çalışması yapmışlardır. Çalışma sonunda iyi dil öğretmeni profilinin sabit tanımı hedeflenmese de var olan çalışma sonuçlarına benzer olarak hassas dengelenmiş sınıf otoritesi, enerji, tolerans, üretkenlik, iyi bir hedef dil bilgisi, sürekli kişisel gelişim, öğrenci özerkliğini geliştirme, iyi iletişim becerileri ve eğitim tecrübesi gibi özelliklerin iyi bir dil öğretmeni ile özdeşleştirildiği görülmüştür. Çalışmada öğretmen adayları ve eğitim planlayıcılarına gelecek için fikir verme amacı güdülmüştür.

Özçelik [32] yaptığı çalışmada; Fransızcayı ikinci yabancı dil (L3) olarak öğrenen üniversite öğrencilerinin dil öğrenme profillerini ortaya koymak ve sınıf, cinsiyet, mezun olunan lise türü, birinci yabancı dil İngilizcenin (L2) düzeyi gibi değişkenlerle bu profil arasındaki ilişkiyi incelemişlerdir. Araştırmaya, Gazi Üniversitesi Gazi Eğitim Fakültesi Fransız Dili ve Eğitimi ABD'de öğrenim gören 160 öğrenci katılmıştır. Araştırmada veri toplama aracı olarak araştırmacı tarafından geliştirilen Yabancı Dil Öğrenme Ölçeği (YDÖÖ) kullanılmıştır. Ölçeğin geçerliliği için alanında uzman iki öğretim üyesinin görüş ve değerlendirmelerinden yararlanılmıştır. Çalışmada faktör analizi metodunu kullanmışlardır, Araştırmada gözle görülür fark yaratan profil; öğrencilerin yabancı dil öğrenme profili üzerinde mezun oldukları lise türünün olmasıdır.

Köktürk [33] liselerin çeşitlenmesi, özel okulların, özel üniversitelerin sayısının artması ile Türkiye de yabancı dil eğitimine talep artmakta olduğunu, bunların arasında İngilizce yabancı dil olarak ağırlıklı bir konumda olduğunu belirtmektedir. Bu bakış açısı ile İngilizce öğretmenlerinin belirli özelliklerde profilinin çizilmesi amaçlanmıştır. Öğretmen profillerinde benzeşmelerin olup olmadığını araştırırken,

demografik özelliklere, motivasyon kavramlarının ifade edilişine, iş tatmin eğilimlerine dikkat edilmiştir.

Özlüer Başer ve ark. [34] 'Makine Öğrenmesi Teknikleriyle Diyabet Hastalığının Sınıflandırılması' isimli çalışmasında Amerika Birleşik Devletlerinde bulunan 130 hastanede ki 7000 kişiye ait verileri toplayarak hastaların diyabet durumlarına göre sınıflandırmayı hedefleyen çalışma yapmışlardır. Sınıflandırma için veri madenciliği tekniklerinden karar ağaçları, k- en yakın komşuluk, lojistik regresyon, naive bayes ve rastgele orman metotları kullanmışlardır. Veri işleme işlemlerinden sonra 2705 kişinin verisi kullanılmış ve 22 bağımsız, 1 tane bağımlı değişken kullanmışlardır. Sınıflandırma algoritmalarının doğru sınıflandırma oranları karşılaştırıldığında rastgele orman yöntemi 4 farklı sınıflandırma ölçütlerine göre en doğru sınıflandırma oranını veren metot olmuştur.

Özcan & Turna [35] gün geçtikçe internet kullanımının artmasından ve özellikle Covid 19 dan sonra e- ticarete olan rağbetten kaynaklı yaptıkları "Karar Ağaçları ile İnternet Alışverişlerinde Tüketiciyi Etkileyen Faktörlerin Analizi isimli çalışmalarında bir tüketiciyi etkileyen faktörleri belirlemeyi amaçlamışlardır. Çalışmada yer alan kişiler 18 yaş ve üzeri kişilerdir. Uygulamada C5.0 VE C&R algoritmalarını kullanılmış, veriler Excel üzerinden hazırlanmış ve SPSS 21 programından yararlanmışlardır. Çalışma sonucunda tüketici yorumları ve ücretsiz kargo olması tüketiciyi etkileyen en önemli faktörler olduğu sonucuna varmışlardır.

Büyükarıkan [36] Finansal Performansa Etki Eden Finansal Değişkenlerin Chaid Karar Ağacıyla Belirlenmesi isimli çalışmasında 21 tane tekstil işletmesinin finansal Performansına etki eden ölçütleri belirlemeyi amaçlamışlardır. Bunun için 20 adet finansal oran kullanmışlardır. Çalışmada IBM SPSS Modeler programı ve Chaıd Algoritması kullanılmıştır, CHAID metoduna göre firmaların performansına etki eden önemli faktörün finansal rantabilite ve sermaye yapısı olduğu belirlenmiştir.

3. MATERYAL VE YÖNTEM

3.1. Veri Madenciliği Tarihçesi

Veri madenciliği kısa bir süre içerisinde hayatımızın her alanında kullanılmaya başlayan, gün geçtikçe hayatımıza dâhil olan ve yakın bir zamanda hayatımıza girmiş gibi görünse de araştırmalar dâhilinde 1700'li yıllarda Bayes Teoremi ve 1800'li yıllarda Regresyon analizi -gibi veri madenciliği uygulamaları ile başlayan tarihi uzak bir geçmişe dayanan bir disiplindir.

Veri madenciliğini tarihi, ilk kez 1946 yılında geliştirilen ve günümüzün kişisel bilgisayarlarının atası olarak kabul edilen ENIAC (Electrical Numerical Integrator And Calculator)'a dayanmaktadır. ENIAC, ABD'li bilim insanları John Mauchly ve J. Presper Eckert tarafından II. Dünya Savaşı sırasında ABD ordusu bünyesinde geliştirilen ilk sayısal bilgisayardır.

1900'lü yılların ortasında bilgisayarlar sayım yapmak amacıyla kullanılmaya başlanmış ve bu kullanım amacı veri madenciliği kavramının çıkması için bir sebep olmuştur. 60'lı yıllarda veri ve bileşenlerine ait kavramlar yavaş yavaş o gün ki teknolojinin içine girmeye başlamıştır. O zamanlarda bilgisayarlarda tarama yapma, istenilen veriye ulaşmanın zor olmayacağı düşünülerek bu konu üzerinde çalışılmıştır. Bu işleme o dönemlerde veri taraması, veri yakalaması gibi isimler verilmiştir. Dönemin sonunda bilim insanları basit mekanizmada sistemler kurmaya başlamışlardır. Bilim insanları birkaç veri madenciliği yöntemlerinin üzerinde çalışmaya ve geliştirmeye başlamışladır. Bu yöntemler üzerinde çalışması ve geliştirilmesi bir takım veri tabanlarını ortaya çıkarmış ve bazı büyük metin depoları sistemde depolanmaya başlanmıştır. Böylelikle veri tabanı yönetme konusunda adımlar atılmıştır.

70'li yıllarda sistem daha da gelişerek bilişsel yönetim veri sistemleri oluşmaya başlamışlardır. Uzmanlar sistemi daha da geliştirerek basit bir makine öğrenmesi geliştirmeye çalışmışlardır. 80'li yıllarda bu yönetim sistemi daha da geliştirilerek

işletmelerin belirli alanlarında kullanılmaya başlanılmıştır ve bir takım veri tabanları oluşturulmaya başlanmıştır.

1980'lerde bağıntılı (relational) veritabanları ve SQL (Select Query Language) yapısal sorgulama dili ile verilerin dinamik ve anlık analiz edilmesine olanak sağlanmıştır [37].

90'lı yıllarda veri miktarı artarak, artan veri nasıl faydalı bir şekilde kullanılmaya başlanır soruları gündeme gelmiştir. 92 senesinde ilk veri madenciliği yazılımı gerçekleştirilmiştir. Bu süre zarfında veri madenciliği ve makine öğrenimi bütünleşmiş çalışarak ilerlemeye başlanmıştır ve makine öğrenmesine yönelik daha özel algoritmalar geliştirilmeye başlanmıştır.

2000'li yıllarda veri madenciliği kavramı gelişerek her alanda uygulanmaya başlanmıştır ve bu alanda çalışmalar artmaya başlanmıştır. 2010'lu yıllarda bu alana ait kavramlar gelişmeye başlanmıştır ve yapay zekâ, öğrenmesi, derin öğrenme vb. kavramlar gelişmiş ve günümüzde her alanda hayatımıza girmeye başlamıştır.

Günümüzde her alanda veri madenciliği kullanılmaktadır ve istatistik ve verinin olduğu her yerde özel veya kamu sektörü fark etmeksizin kullanılmaya devam edecektir. Veri madenciliği yöntemleri ile birçok alanda analiz ve karar verme çalışmaları yapılmaktadır. Teknoloji ve bilgi akışı arttıkça kullanımı artmaya devam edecektir

3.2. Veri Madenciliği Uygulamalarının Gelişimi

Veri madenciliği kavramı çıktığı ilk zamanlarda temel odak yeri çizelge datasıydı. Fakat teknoloji ilerledikçe başka kaynaklarda da madencilik yapma ihtiyacı duyulmuştur. Bunların içerisinde metin madenciliği, görsel madenciliği ve grafik madenciliği yer almaktadır.

3.2.1. Metin madenciliği

Metin madenciliği çalışmalarında veri olarak kabul edilen yapı metindir. Tüm işlem 'metin' üzerinden döner. Metin madenciliğinde metin üzerinden veri elde etme amaçlanmaktadır. Metni sınıflandırma, metin özetleme, metinlerden konu çıkarılması vb. çalışmalar içermektedir.

Metin Madenciliği, işletme dokümanları, müşteri yorumları, web sayfaları ve XML dosyalarını içeren, yapısal olmayan verilerden, önceden bilinmeyen, potansiyel

olarakkullanışlı bilgiyi keşfetme sürecidir. Elde edilen bilgiyle, analiz edilecek olan metinkaynaklarında açık olarak görülmeyen ilişkiler hipotezler veya eğilimler olduğu anlaşılır [38].

Metin Madenciliği, işletme arşivinde veya internet üzerindeki belgelerde bu belgeye benzer belgelerin olup olmadığı elle bir sınıflandırma gerekmeden benzerliği hesaplayabilmektir. Bu genelde otomatik olarak çıkarılan anahtar kelimelerin tekrarı sayesindyapılır [39].

3.3. Veri Madenciliği Uygulama Alanları

Veri madenciliği kısa bir şekilde tanımlamak gerekirse büyük veri yığınlarının analizi yapılarak büyük veri grupları arasından anlamlı sonuç çıkarma işlemidir.

Veri madenciliği belirsiz, net olmayan, kapalı bir potansiyele sahip verilerin açık bir potansiyel haline getirilerek bilginin net bir şekilde görünür halde olmasını sağlar. Faydalı bilgiyi üretme ve bulma çalışması da denebilir.

Bir başka ifade ile veri madenciliği, veri tabanları veya veri ambarlarında yer alan yığın veri içindeki gizliö rüntüleri ve ilişkileri bulmak için istatistiksel algoritmaları ve yapay zekavyöntemlerini kullanan karmaşık bir veri arama yeteneği olarak tanımlanabilir [40].

Günümüzde teknolojinin gelişmesi, kullanımının artması ve buna bağlı olarak veri yığınının fazlalaşması veriler içinden anlamlı bilgi çıkarmayı zorlaştırmıştır. Veri madenciliği de bu karmaşık veriler içinden bir takım istatistik, analiz, kümeleme, sınıflandırma, özetleme işlemlerini kullanarak doğru tespitler çıkarmayı veya gelecekle ilgili tahmin yürütmeyi amaçlayan bir yöntemler çeşididir. Mevcut veri yığınından değerli olan bilgiyi bulma işlemidir. Veri madenciliği elde edilen veriye erişilmesi, temizleme, şeçme işlenmesi, dönüştürülmesi, analiz edilmesi, sunulması ve karar verme işlemlerinden oluşan bir yöntemdir. Özellikle son yıllarda bilgi birikiminin hızlı bir şekilde artması veri madenciliğinin çıkışını hızlandırmış ve çıkış amacının önemi değer kazanmıştır. Veri madenciliği avantajları arasında farklı veriler arasında çalışabilme, eksik veri ile işlem yapabilme, analiz sonucunu sunabilme, veri setinden veri çıkarma ve ekleyebilme işlemleri vardır.

Veri madenciliği hızla gelişen dünyada uygulama alanlarını da geliştirmektedir. Bankacılık sektöründen sağlık sektörüne kadar büyük verinin ve anlamlı sonuçlar

çıkarılabilecek her alanda rahat bir şekilde kullanılan bir yöntemdir. Verinin ve bilginin olduğu her sektörde veri madenciliği tercih edilmektedir.

3.3.1. E-ticaret sektörü

Bu sektörde müşterilerin coğrafik, demografik, kişisel verilerine ulaşarak bir müşteri profili çıkaran çalışmalar uygulamışlardır. Pazar araştırması için müşterilerin satın aldıkları ürünleri takip ederek ürünleri alan müşteriler arasında bir profil çıkarma, ürün satışları arasında bir ağ kurma, e-ticaret sayfalarında önerilebilecek ürünler veya 'bu ürünü alan bunları da aldı' kısmını belirlemek amacıyla bu sektörde veri madenciliği uygulamalarından yararlanılmaktadır. Özellikle metin madenciliğinde e-ticaret sitelerinde müşterilerin yorumlarına göre ürünün beğenme ve beğenmeme oranlarının hesabı yapılmıştır.

3.3.2. Bankacılık sektörü

Veri madenciliği bankacılık sektöründe de oldukça kullanılan literatürde de yer verilen bir sektördür. Bu alanda kredi kartı alan, kredi talebinde bulunan müşterilerin profilini belirleme, kredi kartı, atm dolandırıcılığını tespit etme, müşterilere uygun kampanya belirleme. Müşterilerin ödemelerini düzenli yapıp yapmayacağı ihtimallerini belirlemek vb. faaliyetlerde veri madenciliğinden yararlanılmaktadır. Veri madenciliği metotları ile şirketlerin birtakım istatistikleri yorumlanıp işletme için stratejiler geliştirmektedirler.

3.3.3. Pazarlama sektörü

Pazarlama alanında müşteriler ile ilgili satın alma eylemleri arasındaki ilişkinin belirlenmesi, müşteri profilinin belirlenmesinde, müşterinin değerlendirilmesinde, satış tahmininde bulunulmasında ve bu tarz veri madenciliği ile ilgili pazarlama stratejilerinin geliştirilmesi açısından veri madenciliği bu sektörde de önemli rol oynamaktadır.

3.3.4. Sigortacılık sektörü

Sigortacılık sektöründe veri madenciliği dolandırıcılıklarına karşı tespit, poliçe talep edebilecek müşterilerin tespiti, riskli müşteri gruplarının tespiti vb. durumlarda veri madenciliği yöntemleri kullanılmıştır.

3.3.5. Sağlık sektörü

Sağlık sektöründe yapılan birçok çalışma hastaları elektronik kayıtlarını içeren veriler kullanılarak yapılmaktadır. Bu veriler kullanılarak birden çok tahmin çalışmaları yapılabilmektedir. Örnek vermek gerekirse aynı hastalığa sahip hastaların ortak özeliklerinin belirlenmesi, belli bir hastaneyi tercih eden hastaların demografik özelliklerine göre hasta profilinin belirlenmesi, hastane maliyetleri hakkında tahminde bulunabilme, hastaların hastanede kalış yatış sürelerinin belirlenmesi, hastalık tespitlerinde tarama testlerinin sonuçlarına göre ön teşhis için acil durumlarda hastanın belirtilerine göre risk ve önceliklerin belirlenmesinde bunun benzeri başka bir durumlarda veri madenciliği metotları kullanılmaktadır.

3.3.6. Mühendislik-Bilim alanı

Bu alanlarda fabrikalarda üretim süreci simülasyonu yapılması, kalite kontrol aşamalarında, belirleyici parametreler kullanılarak hava durumu tahmini yapılmasında, deprem verileri kullanılarak deprem öngörüsünde bulunmak vb. durumlar için veri madenciliği bu alanda da kullanılmaktadır.

3.3.7. Emniyet-Güvenlik sektörü

Bu alanda özellikle olası bir suçun engellenmeye çalışılmasında, suçluların işledikleri suçlara ve demografik özelliklerine göre suçlu profili belirleme, suçların bölgesel dağılımı ve buna bağlı olarak suçların önlenmesine yönelik tedbir alınması, suçların istatistiklerine göre raporlanmasında veri madenciliğinden yararlanılmaktadır.

Veri madenciliği farklı sektörlerde kullanım alanları tabloda gösterilmektedir.

Kullanım Alanı	Kullanım Oranı (%)
CRM/ Müşteri Analitiği	32,80
Bankacılık	24,40
Direk Pazarlama	16,10
Kredi Puanlama	15,60
Telekominikasyon	14,40
Dolandırıcılık Tespiti	13,90
Satış	11,70
Sağlık	11,70
Finans	11,10
Bilim	10,60
Reklamcılık	10,60
E- Ticaret	10,00
Sigortacılık	10,00
Web Madenciliği	8,30
Sosyal Ağlar	7,80
İlaç	7,80
Bioteknoloji	7,80

Şekil 3.1. Veri Madenciliği Uygulama Alanları [41].

3.4. Veri Madenciliği Temel Kavramlar

Veri madenciliği; aynı zamanda bilgisayar bilimini, makine öğrenmesini, veritabanı yönetimini, matematiksel algoritmaları ve istatistiği birleştiren disiplinler arası bir alandır [42].

Veri: İngilizcesi data olan ve bu kelimenin de çok fazla kullanıldığı bu kavram bilgisayar çerçevesinde ölçüm sayım araştırma değerlendirme ile bir araya gelen bilginin analize edilebilecek ve değerlendirilebilecek şekilde bir araya gelen yığına veri denir.

Veri işlenmemiş, şekil biçim almamış, düzensiz bir yığın şeklinde ham maddedir.

Enformasyon: Bir sonuca bağlanmak için verinin işlenmiş, anlam kazanmış halidir. Verinin bir amaç doğrultusunda anlamlandırılmış, şekil almış biçimidir.

Bilgi: Enformasyonun bir sonuca yönelik kullanılabilir duruma, sonuç çıkarılabilir hale gelmesidir.

Bilgi sürecine kadar veri 1. Aşama, enformasyon 2. Aşama ve son olarak bilgi 3. Aşama olarak değerlendirilebilir.

Bilgelik: Veri, enformasyon ve bilgiden sonra karar verme aşamasında bilginin sağlıklı bir şekilde kullanılması için geliştirilen anlayış modeli. Kabiliyet, deneyim tarzı özellikler bilgelik unsurudur.

Veri Tabanı: Veri tabanı en kısa tanımıyla bilginin depolandığı yerdir. Günümüzde bilginin hızlı bir şekilde artışı bilginin saklanması gereken bir yer olması gerektiği anlamına gelmektedir. Saklanan bu bilgilerin aynı zamanda kategorize edilen, ayrıştırılan ve istediğimiz zamanda kolaylıkla ulaşabileceğimiz depoya veri tabanı denir. Database olarak da tanımlanabilir. Veri tabanları günümüzde neredeyse her alanda ve sektörde kullanılmaya başlayan depo sistemleridir. Bu depo sistemi bilgiye ulaşmayı kolay ve hızlı bir şekilde gerçekleştirmektedirler. Karmaşık veriyi sınıflandırmak, bir arada tutmak ayrı ayrı dosyalar halinde saklamaktansa veri tabanları bu konuda da avantaj sağlar. Sistemin gelişmesine göre çok eskiye dayalı bilgiye ulaşmamıza da imkân sağlayabilir.

Veri ambarı, iş zekası (BI) faaliyetlerine, özellikle de analitiğe olanak tanımak ve bunları desteklemek üzere tasarlanmış bir veri yönetim sistemidir. Veri ambarları yalnızca sorgulama ve analiz amacıyla kurulur ve çoğu zaman geçmişe ait büyük miktarlarda veri içerir. Bir veri ambarındaki veri genellikle uygulama yazılımlarının günlük dosyaları ve işlem uygulama yazılımları gibi çok çeşitli kaynaklardan elde edilir [43].

Bir veri ambarı, çok sayıda kaynaktan gelen büyük miktardaki verileri merkezi hâle getirir ve birleştirir. Analitik yetenekleri kurumların karar vermeyi geliştirmek için verilerinden değerli iş içgörüleri elde etmelerine olanak tanır. Zaman içinde, veri bilimcileri ve iş analistleri için paha biçilmez bir tarihi kayıt oluşturur. Veri ambarı, bu özellikleri sayesinde bir kuruluşun "tek doğruluk kaynağı" olarak görülebilir.

Veri Bilimi: Ham veriden veriyi değerli bir hale getirecek, analizini soruşturmasını yapacak ve bunlardan anlamlı sonuçlar çıkaracak bir daldır. Veri biliminde asıl amaç karmaşık bilgiden değer yaratacak bilgiyi çıkarma işlemidir ve bu işleme yaparken istatistik, analiz, raporlama, tahmin çalışmalara gibi yöntemler kullanmaktadır. Birtakım kurumların bu veri yığınlarından sağlıklı karar almasına yarayan bir daldır. Bu işlemleri yapan kişilerde veri bilimcisi denir.

Algoritma: Bir problem oluştuğunda veya bir sorunla karşılaşıldığında çözüm için gidilecek yolun adım adım belirtilen işlem adımlarına algoritma denir.

Yapay Zekâ: Yapay zekâ davranış ve düşünce olarak insan beyni ile aynı çalışma prensibine sahip rasyonel ve zeki davranışları olan bir sistem topluluğudur. Yapay

zekâ çok daha geniş bir kapsama sahip olduğu için kesin bir tanım yapmak mümkün değildir.

Makine Öğrenimi: Yazılımın açık bir şekilde kodlanmadan makineye deneyim ve tecrübeler ile öğreten sistemdir. Sisteme vermiş olduğumuz verileri kullanarak benzetim yapan, sonuçlar çıkaran kendini eğiten bir sistemdir.

İşleyiş olarak yapay zekâ ve makine öğrenimi benzer prensipte çalışmaktadır ve kavram olarak birlikte kullanılmaktadır. Fakat aralarında şöyle bir farklılık var ki her makine öğrenimi bir yapay zekâ alt dalı olarak kullanılırken, her yapay zekâ uygulanılan durum bir makine öğrenimi değildir. Yapay zekâ çok daha geniş bir kapsama sahip olduğu için kesin bir tanım yapmak mümkün değildir.

Derin Öğrenme: Veri sayısı arttıkça işlem karışacağından ve daha fazla veriyle daha iyi sonuç alınacağından dolayı sistem makine öğreniminden derin öğrenmeye kaymaktadır. Derin öğrenme makine öğrenmesinin daha üst versiyonudur. Her derin öğrenme bir makine öğrenme algoritmasıdır. Fakat her makine öğrenmesi bir derin öğrenme algoritması değildir [44].

3.5. Veri Madenciliği Süreci

Problemin Tanımlanması, veri madenciliğinin önemli adımlarından biri olarak görülür. Ele alınacak problem anlaşılabilir bir şekilde belirtilirse amaca en uygun veriler veri tabanlarından alınabilir. Bu durumda problemin açık ve net olması sağlıklı veriler kullanarak başarılı bir madencilik işlemi gerçekleştirilmesi olanak sağlayabilir.

Veri Seçimi: Veri madenciliğinde ilk aşamadır. Mevcut probleme çözüm sunabilecek verilerin sistemden çekilme işlemi olarak tanımlanabilir. Sonucunda belli bir analiz elde edileceğinden ve sonucu etkileme durumu olduğunda sistemden verinin doğru ve sağlıklı bir şekilde çekilmesi önemlidir. Verinin kalitesi de bu aşama için önem sarf eden husulardan biri olmaktadır. Fazla ve karışık miktarda verinin tek bir veri ambarı altında toplanması bu aşama için önemlidir. Bu adım fazla zaman alabilir aynı zamanda filtreleme olarak geçebilir. Aynı zamanda verinin alınacağı kaynağın güvenilirliği konusunda net olunması ileride oluşabilecek sorunların önüne geçebilecektir.

Ön İşleme: Bu aşamada veri seçimi kadar önemli ve zaman alabilecek bir aşamadır. İlk aşamada belirlenen işlenmemiş veriler bu aşamada temizlenir ve analizcinin

istediği forma getirilmesi amaçlanır. Bu aşamada toplanan veriden gerekli temizleme işlemi, problem dışı gereksiz bilginin çıkarılması vb. işlemler yapılır. Bu aşamada yapılan sağlıklı işlemlerin sonuçların doğru ve net olması konusunda katkı sağlayacaktır.

Milyonlarca veriden oluşan veri tabanından veri alma işleminden sonra elde olan veride konudan aykırı, fazla ya da tekrarlanmış veri bulunabilir. Veri ön işleme veri madenciliği aşamasında veriyi probleme veya amaca uygun sağlıklı bir hale getirme aşamasıdır.

Veri Temizleme: Eksik özellikteki veriyi tamamlama, yanlış verileri düzeltme, gereksiz bilgiyi saptama ve temizleme işlemleridir.

Ayrıca veride bulunan eksik veya kayıp bilgiler de gözden geçirilmelidir. Örneğin bir veri tabanında yer alan kişilerin medeni hali belirliyken, bazı kayıtlarda bu bilgi eksik olabilir veya bu kayıt hiç girilmemiş olabilir. Bu durumdaki eksiklik "kayıp veriler" olarak tanımlanabilir. Bunun haricinde, bazı kayıtlarda aşırı uç değerler (outlier) veya yanlış girilmiş değerler olabilir. Bu tür bilgilere de gürültü (noise) adı verilmektedir [45].

Veri Birleştirme: Farklı veri tabanlarından alınan verinin birleştirilmesi işlemidir. Farklı veri kaynaklarından alınan veriler tak bir çatı altında birleştirildikten sonra da hatalar ortaya çıkabilir. Farklı tablolardan gelen veriler birleştirildiği zaman fazla veri ortaya çıkmaktadır. Bu da veri kümesinin gereksiz fazla olmasına ve veri madenciliği işleminin yavaş olmasına sebep olmaktadır.

Veri Dönüştürme: Veri madenciliğinde kullanılan metoda göre veri üzerinde bir takım dönüştürme işlemi yapılır. Cevabı evet/ hayır olarak alınan bir soru veri üzerinde 0/1 olarak gruplandırılır. Başka bir örnekte belli bir sayı aralığındaki değerler aynı şekilde yapılacak madene göre düşük/orta/yüksek şeklinde gruplandırılabilir.

Veri İndirgeme: Kaynaktan alınan veriler orijinal yapıları nedeniyle madencilik işlemi esnasında yöntem için uyarlanamayabilir. Özellikle çok büyük sayı değerinde ve çok fazla farklı sayıda sayı verileri olduğu zaman değerler daha küçük ve az sayıda değer aralığında atanır. Bu sebepten veriler üzerinden bir takım dönüştürme işlemi yapılmaktadır. Çözüm için veri düzeltme, biriktirme, genelleme, normalizasyon, nitelik oluşturma işlemleri yapılmaktadır. Bazı kaynaklarda veri dönüşümü veri indirgeme olarak da adlandırılmaktadır.

Normalizasyon: Verileri 0-1 arasında yeniden ölçeklendirme işlemi yapan işlemdir.

Min-max normalizasyon: Bu normalleştirme tekniği değerler, başka bir sayı değeri aralığına uyarlanarak dönüştürülür. Genellikle denklem 3.1 kullanılır. Bu sayı değeri aralığı olarak 0-1 sayı aralığı kullanılır.

$$X = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{3.1}$$

Bu normalizasyon işleminde en küçük değer 0, en büyük değer 1 olarak alınarak bu formül ile diğer verilerinde bu sayı aralığında ataması yapılır.

Z-score normalizasyon: Bu normalizasyon türünde verilerin ortalaması ve standart sapması kullanılarak her bir değişken için normalleştirme işlemi yapılır. Bu işlemler için denklem 3.2, 3.3 ve 3.4 kullanılır.

Ortalama:

$$\mu = \frac{1}{N}\sum_{i=1}^{N}(X_i) \tag{3.2}$$

Standart Sapma:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_{i-}\mu)^2} \tag{3.3}$$

Z-Score:

$$\mathrm{Z}=\frac{x-\mu}{\sigma} \tag{3.4}$$

Ondalık Normalizasyonu: Veri kümesi içindeki değerlerin ondalık kısmına göre işlem yapılmasıdır.

Veri Madenciliği: Veri bu aşamada madenciliği hazır bir haldedir. Probleme uygun veri madenciliği metodu seçilmelidir. Elimizde bulunan veriden en verimli sonucu alabilmek için bu noktada kullanılacak model çok önemlidir. Modelin doğru belirlenmesi analizden sağlıklı sonuçlar alabilmemiz için önemli bir faktördür. Modelin doğru kurulmaması durumunda veriler arasındaki bağlantı sağlıklı bir şekilde kurulamaz ve sonucun başarılı olma olasılığı da bu nokta da düşer. Burada ki en önemli husus probleme yönelik ve problemin ihtiyaçlarını karşılayacak metodun bulunmasıdır.

Çalışmanın amacına göre veri madenciliği metotlarından bir veya birkaçı uygulanır.

Veri madenciliği model kurulum aşamasında kullanılan modellere göre Tahmin edici modeller ve tanımlayıcı modeller olmak üzere ikiye ayrılmaktadır. Veri madenciliği işlevlerine göre üç sınıfa ayrılır. Bu sınıflar sınıflandırma, kümeleme ve birliktelik kurallarıdır.

Yorumlama ve Değerlendirme: Veriler üzerinden veri madenciliği uygulandıktan sonra çıkan sonuçlar analiz edilir ve yorumlanır. Eğer birden çok yöntem kullanılmış ise sonuçlar karşılaştırılır.

3.6. Veri Madenciliği Metotları

Veri madenciliği tahmin edici ve tanımlayıcı modeller olmak üzere iki başlık altında incelenir.

Tablo 3.1. Veri Madenciliği Metotları.

Tahmin Edici Yöntemler	**Tanımlayıcı Yöntemler**
Sınıflandırma	Kümeleme
Karar Ağaçları	Birliktelik Analizi
En Yakın Komşu Algoritması	Sıralı Dizi Analizi
Yapay Sinir Ağları	Özetleme
Bayes Sınıflandırması	Tanımsal İstatistik
Zaman Serisi Analizi	
Karar Destek Makineleri	
Diğer Yöntemler	

3.6.1. Tahmin edici yöntemler

Tahmin edici modellerde daha önce sonucunu bildiğimiz veri setini kullanarak sistemin oluşturduğu modele göre sonucunu bilmediğimiz verilerin tahmin çalışması yapılmaktadır. Sınıflandırma modelleri farklı sektörlerde de tercih edildiğinden en çok kullanılan veri madenciliği metotlarından biridir. Sınıflandırma Modelleri tahmin edici modeller arasında yer almaktadır. Veri seti içerisindeki bağımsız değişkenleri alarak öngörü yapma yeteneğine sahiptir.

3.6.1.1. Karar ağaçları

Veri madenciliği metotlarından karar ağacı sınıflandırma grubundan oldukça yaygın kullanılan verilmek istenen kararın olası durumlarını bir bütün olarak görmemizi sağlayan, tepeden aşağıya doğru inen bir ağaç yapısına benzetilerek kullanılan bir veri madenciliği yöntemidir.

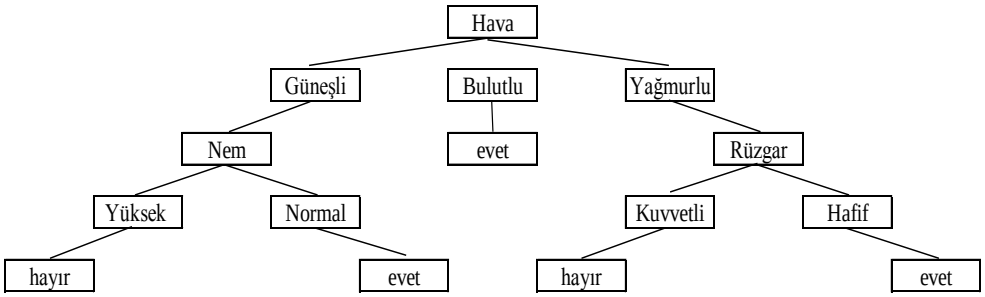
Karar ağacı metodu anlaması ve yorumlaması kolay olduğundan, görselleştirilme özelliğinden, kısa sürede sonuç vermesinden, hedefin türüne göre algoritma seçebilme özelliğinden dolayı tercih edilen bir yapıdır.

Karar ağaçları problemin çözümüne giden yolda diğer ihtimalleri de bir bütün halinde görmemizi sağlayan grafiksel bir tarzı olan bir metottur. Büyük veri yığınlarını küçük kümeler halinde bölerek işlem yapan bir yapıdır.

Karar verici, türlü seçeneklerin gerçekleşmesinin belirli ya da belirsiz olduğu bir problemle ilgili en iyi karara ulaşmak amacıyla, bazı işlemlerin yerine getirilmesi için birtakım yöntemlere veya araçlara gereksinim duyar. Seçenek sayısının fazla olduğu ve/veya ardışık aşamalarda karar almanın söz konusu olduğu problemlerin analizi, modellerin kurulması ve çözümlenmesi işlemlerinde karar vericiler bu araçlardan birisi olan "Karar Ağacı Analizi"ni kullanabilirler [46].

Bir karar ağacı, kök düğüm adı verilen bir değişkenden başlayan ağaç benzeri bir yapıda hiyerarşik bir ilişki grubudur. Bu kök düğüm, kök düğümün ayrı sınıflarını veya düğümün ölçeği boyunca belirli aralıkları temsil eden çok sayıda dalda iki bölüme ayrılır. Her bölüntüde, bölünen değişkenin sınıfları veya aralığı bakımından yanıtı olan bir soru sorulmaktadır. Bu soru örneğin, "erkek mi kadın mı?" olabilir. Bunun gibi sorular, ikiye bölünmüş karar ağacı oluşturmak için kullanılır. Karar ağaçları birden çok bölme ile de oluşturulabilir. Her bir bölünmede sorulan sorular, sonuçta ortaya çıkan vakaların bölünmelerde ne kadar üniform olması gerektiğini yansıtan bazı belirsizlik ölçüleri açısından tanımlanır. Her dal, diğer değişkenlerin sınıfları veya aralıkları kullanılarak daha da bölünür. Her bölüntüde bölünen düğüme ana düğüm, bölünmüş olduğu düğümlere de alt düğüm adı verilir. Bu işlem, kesme kuralı gerçekleşinceye kadar devam eder [47].

Aşağıdaki şekil 3.2 en basit haliyle karar ağacı örneği verilmiştir.



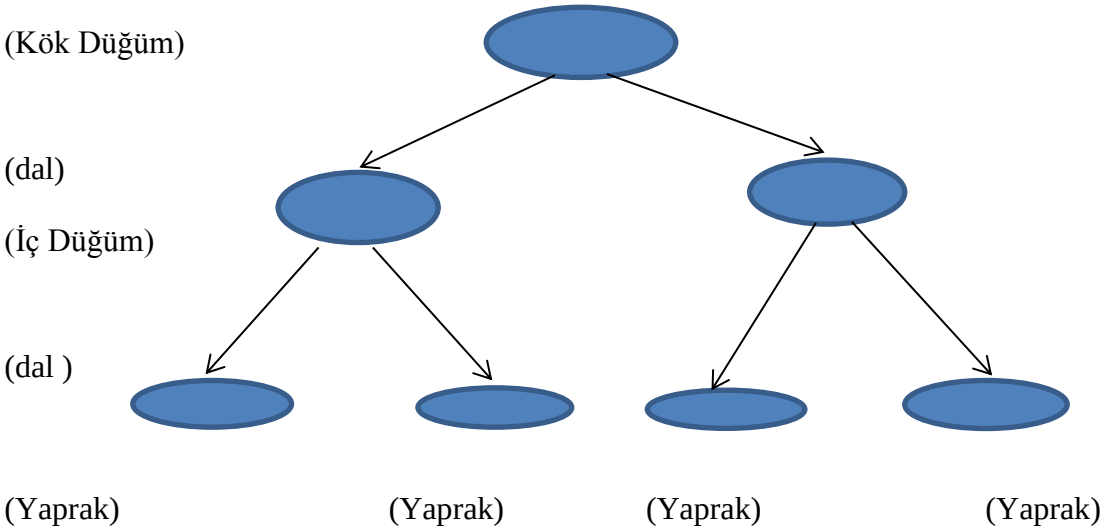
Şekil 3.2. Karar Ağacı Örneği.

3.6.1.2. Karar ağaçlarında dallanma kriterleri

Karar ağacı metodunda en önemli noktalardan biri dalların ayrılmalarının neye göre yapılacağıdır. ID3 ve C4.5 algoritmaları entropiye dayalı, CART algoritması Twoing ve Gini sınıflandırma ve regresyon ağaçlarına dayalıdır.

Karar ağaçları oluşturmada en önemli adım ağaç dallanması için gerekli kriterlerin belirlenmesidir. Bu sorunu çözmek amacıyla çeşitli yaklaşımlar geliştirilmiştir. Bunların en önemlileri arasında 'bilgi kazancı' ve 'bilgi kazancı oranı', 'gini-indeksi', 'towing-rule' ve 'kare olasılık tablosu istatistiği' yaklaşımları sayılabilir. Bu yaklaşımları kullanarak ID3, C4.5, C5.0, CART, CHAID ve QUEST gibi sınıflandırma algoritmaları geliştirilmiştir [48].

Aşağıdaki şekil 3.3 de karar ağacı düğüm yapısı örneği verilmiştir.



Şekil 3.3. Karar Ağacı Düğüm Yapısı.

Entropi: Bir sistemde belirsizliği ölçen değere entropi denir.

Kısa bir tanımıyla entropinin düşük olduğu yerde düzen vardır, entropi değerinin fazla olduğu küme içerisinde daha fazla düzensizlik vardır. Entropi küme içerisindeki homojenlik ile ilgilenir.

Karar ağacı algoritmalarından yaygın olanlar aşağıda açıklanmıştır.

ID3 Algoritması: Sydney Üniversitesi'nde araştırmacı olan J. Ross Quinlan tarafından geliştirilmiştir. Veri setinde verilen örnekler arasında farklı değişkeni makine öğrenmesi ve bilişim teknolojisi aracılığıyla bularak, işlem esnasında entropiden yararlanan bir algoritmadır. Entropi, verileri birbirinden ayıran farklılıklarıdır. Entropi, sistemdeki belirsizliği tespit etmekte ölçüttür ve bir alanın entropi ölçüsü yüksek olması, mevcut olanın belirsizliğini arttırmaktadır. Bu nedenle, doğru sonuca varmak adına karar ağacının kökünde entropi ölçüsü en az belirsizlik ve kararsızlık içeren değer kullanılmaktadır [49].

ID3 Algoritması karar ağacı algoritmalarından en basit yapıda olan algoritmadır. Algoritmanın temel çalışma yapısı uygulaması kolay ve anlaşılması rahat bir yapıdadır. Algoritmanın çalışma yapısında entropiden ve bilgi kazanımından faydalanılmaktadır. Entropi sistemde ki düzensizlik üzerine çalışır. Sistemde ki karar noktaları entropi üzerinden değerlendirilir.

C4.5 Algoritması: C4.5 Algoritması ID3 Algoritmasının gelişmiş versiyonudur. C4.5 algoritmasını ID3 Algoritmasından ayıran özellik numaralaştırılmış değerler kategorilendirmiş değerler olarak ayrılır. ID3 algoritmasında bilgi kazançları hesaplanır ona göre işleme devam edilir. C4.5 algoritmasında ise en yüksek değeri verebilecek bir eşik değer hesaplanır.

C4.5 algoritmasında kayıp veriye ilişkin tüm değerler veri tabanından çıkartılır ama ID3 algoritması bu konuda işlem yapamaz. C4.5 algoritması ID3 algoritmasına göre daha hızlı ve daha özenli işlem yapıyor olmasıdır.

C4.5 algoritması entropi kavramını baz alarak işlem yapar. Entropi değeri hesaplaması denklem 3.5 de gösterilmiştir.

$$\text{Entropi (S)} = \sum_{i=1}^{n} -pi \log_2(pi) \tag{3.5}$$

P=(M1, M2, …Mn) olmak üzere n adet eleman olduğunu farz edelim. Bu Mi elemanının sistemde bulunma olasılığı Pi' dir.

P=(P1,P2,…..Pn) olasılıklarına sahip sistemin entropisi yukarıdaki formüle göre bulunur.

Entropi hesaplanmasını bir örnekle ele alalım;

Örn: 8 eleman içeren bir X kümesi ele alalım.

X=(kırmızı, mavi, kırmızı, mavi, mavi, mavi, mavi, mavi)

Kırmızı ve mavi için olasılık değerleri

P(kırmızı)= 2/8 P(mavi) =6/8

Bu sınıf kümesi için ortalama entropi değeri;

Entropi (S) = - $(\frac{2}{8}\log_2\frac{2}{8} + \frac{6}{8}\log_2\frac{6}{8})$ = 0,811

Karar ağaçları yöntemlerinde baskın (ayırt edici) niteliği belirlemek için her niteliğin bilgi kazancı hesaplanır. Bir A özelliğinin S örneği için bilgi kazancı Denklem 3.6 ’e göre yapılmaktadır [49].

$$\text{Kazanc(A, S)} = \text{Entropi (S)} - \sum_{j=1}^{n} \frac{S_j}{S} Entropi\ (S_j) \qquad (3.6)$$

Yukarıdaki denklem 5.6 temel alınarak belirlenen her özellik için bilgi kazancı hesaplanır. Bilgi kazancı enyüksek olan özellik kök olarak tanımlanır. Bu işlemler her düğüm için;
-Örneklerin aynı sınıfa ait olması

-Örneklerin bölünebilecek özelliklere ayrışamaması

-Tüm özelliklerin uygun sınıfla temsil edilmesi gibi durumlardan biri oluşuncaya kadar devam eder [50,51].

Bütün kazanç değerleri belirlendikten sonra en yüksek kazanç değerine sahip nitelik dallanma düğümü olarak belirlenir.

CART Algoritması: Cart Algoritması ismini ‘Classification and Regression Trees’ İngilizce isminden almaktadır. Türkçe’ ye Sınıflama ve Regresyon Ağaçları Tekniği olarak geçmektedir.

1984 senesinde algoritma Breiman, Friedman, Olshen ve Stone sayseinde geliştirilmiştir. CART algoritmasında nitelik olarak belirtilen bağımsız değişkenler ağacın farklı seviyelerinde birden fazla görülebilmektedir. CART algoritması, ID3 ve C4.5 ve C5.0 algoritmaları gibi en iyi dallara ayırma kriterini belirlerken entropiden

yararlanmaktadır. Ancak bunun için bu algoritmalardan başka iki yöntem kullanmaktadır [52].

CART algoritması her adımda ilgili grubun, kendisinden daha fazla homojen 2 alt gruba ayrılmasına yaramaktadır. Yani her dal ikili alt dallara bölünerek büyümektedir. Bu ayırma da bağımlı değişken türü kategorik ise twoing**,** gini indeksi, sürekli değişkense en küçük kareler sapması hesaplamasına göre yapılmaktadır. CHAID algoritması gibi hem sınıflama hem de regresyon amacıyla kullanılmaktadır [53]. Twoing ve Gini algoritmaları sınıflama adı altında geçmektedir. En küçük kareler sapması ise regresyon adı altında geçmektedir. Bu metot da ikili seçimler ile homojen sınıflar oluşturarak işleme devam edilmektedir. Bu algoritma kayıp değerler yerine o değerin yerini tutabilecek bir değer atayabilmektedir.

Twoing Algoritması, : Bu algoritmada tüm veri kümesindeki öznitelikler öncelikle sağ ve sol olmak üzere 2 gruba ayrılır. Örnek vermek gerekirse bir bir kümede 'YAŞ' niteliğinin 3 tane belirleyici faktör bulunsun. Bu faktörler 'genç, orta, yaşlı' olarak belirlensin. Twoing algoritması bu niteliği ayırırsa ortaya alttaki tablo 3.1 çıkmaktadır.

Tablo 3.2. Twing Algoritması Örneği.

Bölünme	SOL	SAĞ
1	Yaş=genç	Yaş= {orta, yaşlı}
2	Yaş=orta	Yaş={genç, yaşlı}
3	Yaş=yaşlı	Yaş={genç, orta}

Bölünmelerden sonra Psol, Psağ, P(j/Tsol), P(j/Tsağ) olasılık değerleri bulunur. Bulunan hesaplamalardan sonra aşağıdaki denklem 3.7 ile uygunluk değeri bulunur.

$$\Phi(s/t)=2.Psol.\ Psağ.\sum_{j=1}^{n}|Pj(Tsol)-Pj(Tsağ)| \quad (3.7)$$

Φ değerleri hesaplandıktan sonra uygunluk değeri en yüksek olan seçilir. En yüksek değer bize dallanmanın olacağı niteliği belirler. Bu şekilde döngü her alt küme için geçerli olur.

Gini Algoritması: Gini algoritmasında veri kümesinde nitelikle 2'li olacak şekilde ayrılır. Her nitelik için bu bölünme yapıldıktan sonra sağ ve sol taraf için Gini değerleri bulunur. Aşağıdaki formülde Li soldaki belirlenen kategorideki örnek sayısı, Ri sağdaki belirlenen kategorideki örnek sayısı, Tsol sol taraftaki toplam örnek sayısı, Tsağ sağ taraftaki toplam örnek sayısını belirtir. Denklem 3.8 ve 3.9 da gösterilmiştir.

$$\text{GiniSol} = 1 - \sum_{i=1}^{k} \left(\frac{Li}{|Tsol|}\right)^2 \tag{3.8}$$

$$\text{GiniSağ} = 1 - \sum_{i=1}^{k} \left(\frac{Ri}{|Tsağ|}\right)^2 \tag{3.9}$$

Her bir nitelik için GiniSol ve GiniSağ değerleri bulunduktan sonra her bir nitelik için Gini değeri hesaplanır. Bu değerin hesaplanması denklem 3.10 da gösterilmiştir.

$$\text{GiniJ} = \frac{1}{n}(|Tsol|.GiniSol + |Tsag|.GiniSag) \tag{3.10}$$

Gini değeri en küçük olan nitelik seçilir ve bu nitelikten karar ağacı dallanmaya başlar.

CHAID Algoritması: CHAID algoritması açılımı Chi-squared Automatic Interaction Detector'den gelmektedir. Türkçe'ye Otomatik Ki-Kare Etkileşim Belirleme Analizi olarak geçmektedir. Ki-Kare olarak geçmesinin sebebi ise algoritmanın arka planında çapraz tablolar ile işlemlerin olmasıdır. CHAID algoritması Kass tarafından 1980 senesinde ortaya atılmıştır CHAID algoritması dallanma aşamasında bağımlı değişkenin kategorisine göre farklılık gösterir. Eğer bağımlı değişken kategorik ise ki-kare, sürekli ise f-testi kullanılarak işleme devam etmektedir. Algoritma işleme devam ederken homojen değerleri bir araya toplayarak, geriye kalanları ise heterojen olarak değerlendirerek devam eder. Devamında dallanmaya başlayacak en uygun değişkeni tespit eder ve işlem bu şekilde devam eder. Algoritmanın asıl amaçlarından biri veriyi daha homojen bir şekilde alt dallara ayırmaktır ve bu şekilde daha tutarlı sonuçlar elde etmektir.

Bu algoritmanın farklı çeşit bağımlı değişkenlerle çalışabilmesi, tercih edilebilir bir algoritma olmasını da sağlamıştır. Aynı zamanda ikiden fazla dala ayrılması CHAID Algoritması için bir avantaj olarak yer almaktadır. Araştırmalar neticesinde en fazla pazarlama alanında kullanıldığı saptanmıştır.

CART VE CHAID Algoritmaları Karşılaştırması: Bu iki algoritma temelde birbirine benzemektedir. CHAID algoritması işlem sırasında ki-kare testi kullanmaktadır. CART algoritmasında entropiden yararlanılmaktadır. CART

algoritmasında ikili ağaç yapısı yer almaktadır, CHAID algoritmasında çoklu ağaç yapısı bulunmaktadır.

QUEST Algoritması: QUEST algoritması 1997 senesinde geliştirilmiştir. Bu algoritma birtakım özelliklerinden CART ve CHAID algoritmalarından ayrılmaktadır ve daha hızlı sonuç vermektedir. Bu kararağacında bağımlı değişken kategorik olmak durumundadır. Bağmsız değişken kategorik ise ki-kare, sürekli ise f testi uygulanmaktadır.

3.6.1.3. K-en yakın komşu algoritması

Literatürde K-NN olarak geçmektedir ve açılımı K-Nearest Neighbor şeklindedir. Literatürde ki bilgilere göre 1950'li yıllarda adı geçmeye başlanmıştır. K en yakın komşu algoritması sınıflandırma modellerinde en çok kullanılan modellerden biridir. Hem sınıflandırma hem de regresyon modellerinde kullanılmaktadır. Amaç verilerin hangi kategoriye ait olduğu bilinen veri kümesine yeni katılacak bir değerin hangi kategoriye ait olduğunu saptamaktır.

Bu algoritmada tüm veriler bir çatı altında toplanır ve yeni veri için atanacak en yakın sınıf seçilir ve belirsiz veriye en yakın k sınıflarını belirler.

En kısa tanımıyla belli bir sınıfa ait olmayan veri örneklemde ki diğer veriler ile kıyaslanır ve bir uzaklık belirlenir. Hesaplanan uzaklığa göre belli bir sınıfa ait olmayan veri için en uygun sınıf belirlenir. Bu algoritmanın şöyle bir avantajı vardır; çok hızlı değişen sistemlerde veriler hızlı bir şekilde gelip hızla bir şekilde değişebildiğinden bu değişime ayak uydurması daha hızlı bir şekilde olmaktadır. Ana amaç benzer verilerin aynı sınıfta toplanmasıdır.

Algoritma için en kullanılan uzaklık hesaplama yöntemleri aşağıdaki gibidir;

Euclidean (Öklidyen) Uzaklık hesaplaması,

Manhattan Uzaklık Hesaplaması,

Chebyshev Uzaklık Hesaplaması,

Hamming Uzaklık Hesaplaması,

Minkowski Uzaklık Hesaplaması,

Mahalanobis Uzaklık Hesaplaması,

Haversiene Uzaklık Hesaplaması,

Levenshtein Uzaklık Hesaplaması,

Sørensen-Dice Uzaklık Hesaplaması,

Jacckard Uzaklık Hesaplaması.

En çok tercih edilen uzaklık ölçütleri Euclidean başta olmak üzere, Manhattan ve Minkowski uzaklık ölçütleridir.

Euclidean Uzaklık Ölçütü: Euclidean uzaklığı, iki nokta arasında, $x1$= ($x11$, $x12$ … $x1n$) ve $x2$= ($x21$, $x22$ ..$x2n$)

Olmak üzere denklem 3.11 ile

$$d(i,j) = \sqrt{\sum_{k=1}^{n}(X_{ik} - X_{jk})^2} \quad (3.11)$$

Manhattan Uzaklık Ölçütü denklem 3.12 ile

$$d(i,j) = \sum_{i=1}^{n}\left|X_{ik} - X_{jk}\right| \quad (3.12)$$

Minkowski Uzaklık Ölçütü

$$d(i,j) = \left[\sum_{i=1}^{k}\left|X_{ik} - X_{jk}\right|^{q}\right]^{1/q} \quad (3.13)$$

Bu algoritmanın adımları şu şekildedir;

-Öncelikle bir parametre seçilir ve seçilen parametre k olarak adlandırılır.

-Daha sonra belirlenen noktanın diğer noktalara olan uzaklığı uzaklık ölçütleri ile bulunur.

-Belirlenen noktaya diğer tüm noktaların uzaklıkları bulunduktan sonra uzaklıklar küçükten büyüğe sıralanır ve en küçük uzaklıklar k parametresi kadar seçilir.

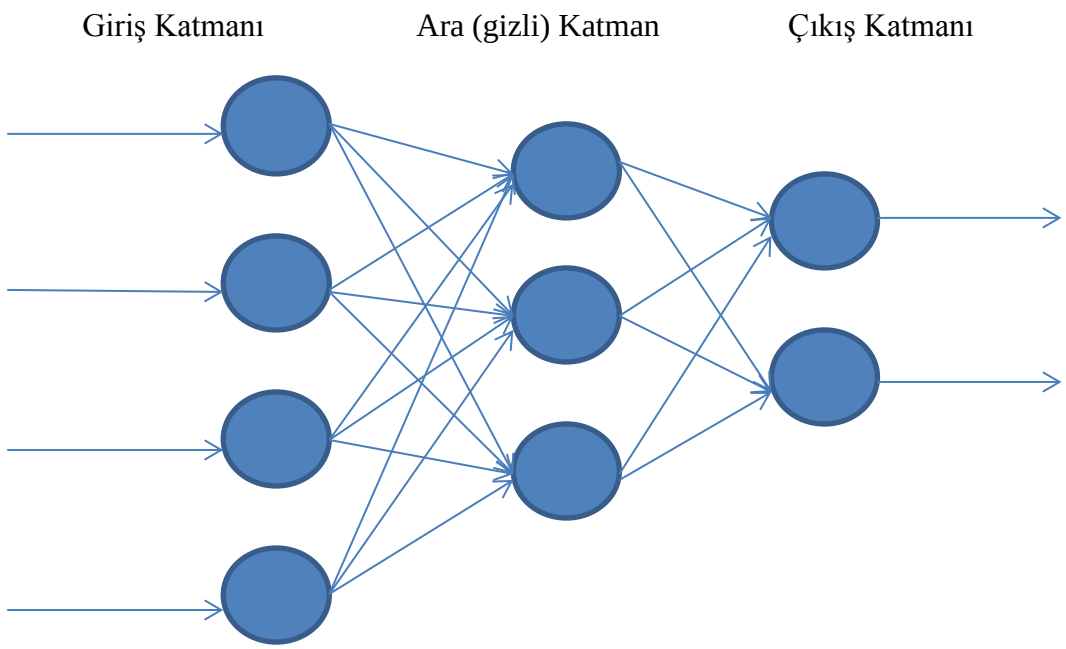
-K parametresi kadar seçilen noktalarında sınıfları belirlendikten sonra hangi sınıf daha çok baskın ise belirlenen nokta o sınıfa atanır.

3.6.1.4 Yapay sinir ağları

Teknolojinin de, teknolojide kullanılan bilgisayarlarında vazgeçilmez bir yere sahip olduğu bellidir. Günümüzde artık her alanda kullanılan bilgisayarların insan beyni mekanizması gibi karar verme ve öğrenme yöntemlerini kazanması ile kullanım

alanları gün geçtikçe genişlemiştir. Matematiksel olarak ifade edilemeyen ve insanlar tarafından çözülmesi zor olan problemlerin çözümünde yapay zekâ yöntemleri kullanılmaktadır. Yapay zekâ yöntemlerinin en temel niteliği, olaylara ve problemlere çözümler üretirken örneklerden, tecrübeden, benzetmelerden yararlanarak öğrenme gerçekleştirme ve karar verebilme yeteneklerinin olmasıdır. Makine öğrenimi için en popüler metotlardan biri de yapay sinir ağlarıdır. Yapay sinir ağları, geleneksel hesaplama yöntemleri ile çözülemeyen problemlerin çözümünde yaygın olarak kullanılmaktadır. Öğrenme, genelleme, doğrusal olmama, hata toleransı, uyum, paralellik gibi özellikleri olan ysa; görüntü ve sinyal işleme, hastalık tahmini, gibi tıbbi uygulamalarda; mühendislik, üretim, finans, optimizasyon, sınıflandırma gibi çok farklı uygulama alanlarında yer almaktadır.

Yapay Sinir Ağları Yapısı: Yapay sinir ağları birden fazla yapay sinir hücresinden meydana gelir. Bu ağın yapısı temel olarak üç ana katmandan oluşur. Bu katmanlar girdi katmanı, ara katman ve çıktı katmanıdır. Çok katmanlı ileri beslemeli yapay sinir ağı yapısı örneği şekil 3.4 de verilmiştir.



Şekil 3.4. Yapay Sinir Ağı Yapısı.

Yapay Sinir Ağları öğrenme şekillerine göre;
-Eğitimli Öğrenme-Eğitimsiz Öğrenme
-Yarı Eğitimli ÖğrenmeYapay Sinir Ağları öğrenme kurallarına göre;
-Geri Yayılım Algoritması
-Hebb Kuralı
-Hopfield Kuralı
-Kohonen Kuralı
-Delta Kuralı
Ağ yapılarına göre Yapay Sinir Ağları;
-İleri Beslemeli Yapay Sinir Ağı
-Geri Beslemeli Yapay Sinir Ağı
Öğrenme kuralları ve ağ yapılarına göre en sık kullanılanlar Geri yayılım algoritması ve ileri beslemeli ağ yapısı olduğundan bu iki yapıdan bahsedilmiştir.

Geri Yayılım Algoritması: Bu algoritmanın temel prensibi oluşan hatanın geriye doğru aktarılmasıdır. Gerçek değerler ve hesaplanan değerler arasındaki hata geriye doğru bir hareketle gizli katmandaki nöronlara aktarılarak ağırlıklar değiştirilir. Bu algoritmanın avantajında modelin revize edilmesi vardır.

İleri Beslemeli Yapay Sinir Ağı: Bu yapay sinir ağında adımlar giriş katmanından çıkış katmanına doğru ilerler. Giriş katmanına aktarılan veriler aynı şekilde gizli katmana gönderilir. Burada da bir değişikliğe uğramadan çıkış katmanına iletilir.

3.6.1.5 Bayes sınıflandırması

Naive Bayes sınıflandırıcısı adını İngiliz matematikçi Thomas Bayes'ten alır. Bayes sınıflandırıcıları istatiksel sınıflandırma teknikleri arasında yer alır. İstatiksel Bayes teoremine dayanır, tahmin edici modeldir ve basit uygulanabilir bir yöntemdir. Naive Bayes, hedef değişkenle bağımsız değişkenler arasındaki ilişkiyi analiz eden tahminci ve tanımlayıcı bir sınıflama algoritmasıdır. . Naive Bayes, sürekli veri ile çalışmaz. Bu nedenle sürekli değerleri içeren bağımlı ya da bağımsız değişkenler kategorik hale getirilmelidir. Örneğin; bağımsız değişkenlerden biri yaş ise, sürekli değerler “<20” “21-30”, “31-40” gibi yaş aralıklarına dönüştürülmelidir [54].

Naive Bayes sınıflandırma algoritması esasında bayes teoremini baz alarak işlem yapar. İşlemler sonucu hesapladığı değer ile verilerin hangi sınıfa yüzde kaç olasılıkla ait olduğu ile ilgilenir. Bu olasılığın hesaplanabilmesi için aşağıda ki denklem 3.14 kullanılır.

$$P(A/B) = \frac{P(A)P(B/A)}{P(B)} \quad (3.14)$$

P(A/B)=B durumunun gerçekleştiğinde A olayının gerçekleşme olasılığı

P(A)=A durumunun gerçekleşme olasılığı

P(B/A)=A durumunun gerçekleştiğinde B olayının gerçekleşme olasılığı

P(B)=B durumunun gerçekleşme olasılığı

Zaman Serisi Analizi: Zaman serisi analizinin nerelerde kullanıldığını anlamak için zaman serisinin ne olduğu iyi bilinmelidir. Her seri zaman serisi değildir, bir serinin zaman serisi olabilmesi için zamana bağlı bir durum olmalıdır. Örneğin borsa değeri bir zaman serisidir, borsa değeri hesaplanırken bir önceki günün kapanış değeri bir sonraki günün değerini etkilemektedir [55].

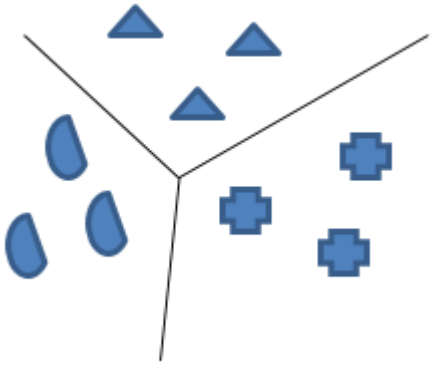
Bir seriyi zaman serisi olarak değerlendirebilmemiz için sistemin içinde zamana bağlı durum gerçekleşmesi gerekmektedir. Eğer durumun içinde zaman varsa o zaman serisi olarak adlandırabiliriz. Örneğin bir mağazaya gelen müşteriler hafta sonları fazla olurken, hafta içi az olabilmektedir bu noktada zamana bağlı bir durum vardır.

3.6.2. Tanımlayıcı yöntemler

Tanımlayıcı modeller karar verme aşamalarında yardımcı olmak amacıyla kullanılan modellerdir. Mevcut veri setinde veriler arasında ilişkilerin tanımlanmasına yardımcı olur.

3.6.2.1 Kümeleme analizi

Kümeleme analizi isminden de anlaşılacağı üzere benzer niteliklere sahip verilerin aynı veri grubu içerisinde toplanmasıdır. Birbirinden farklı nitelikte ki veri yığınlarını homojen gruplara ayıran metottur.



Şekil 3.5. Kümeleme Analizi Örneği.

3.6.2.2 Birliktelik analizi

Araştırmalara göre birliktelik analizi veri madenciliğinde kullanılan ilk metotlardan biridir. İsminden de anlaşılacağı üzere birliktelik analizi veri tabanından geçmiş verileri temel alarak birlikte olan olayların tespitini yapmaktadır. Bu yaklaşımı tüm veri yığını içerisinden birlikte gerçekleşen durumları saptayıp ileriye yönelik iyileştirme çalışmaları yapmayı amaçlamaktadır.

Birliktelik kurallarının ilk kullanımı perakendecilik sektöründe olmuştur. Perakendecilik sektöründe gerçekleşen işlem verilerinin analiziyle ürün yerleşimi ve satış artımına yönelik kullanılmıştır. Örneğin; Amerikan perakende mağaza zinciri Wal-Mart'ın yaptığı araştırmaya göre bebek bezi ve bira arasında güçlü bir ilişki vardır. Yapılan analizler, Cuma günü saat 17:00 ile 19:00 saatleri arasında bebek bezi almaya gelen müşterilerin çoğunun bira da aldığını göstermektedir. Wal-Mart yetkilileri bu sonuca dayanarak bebek bezi ve bira reyonlarını yan yana getirmiş ve bebek bezi alıp bira almayan müşterilerin bile bira aldığı yapılan analizler sonucunda ortaya çıkmıştır [56].

Birliktelik analizi için kullanılan algoritmalar; Apriori, AIS, STM, CARMA, DIC ve bu algoritmalardan başka birden çok algoritma kullanılmaktadır. Fakat içerisinde en çok tercih edilen Apriori Algoritmasıydı.

4. UYGULAMA VE ARAŞTIRMA BULGULARI

4.1. Problemin Tanımı

Bu çalışmada 2021 yılında faaliyete başlayan bir yurt dışı eğitim firmasının kayıt altına aldıkları müşterilerin verilerine göre bir müşteri profili çıkarması amaçlanmıştır. Çalışma sonunda müşterilerin tanımlanması, hedef müşteri grubunun belirlenmesi, özel teklif sunulabilecek müşterinin profilinin belirlenmesi, gerekli durumlarda müşteri portföyünün değiştirilmesi ve en çok müşteri aldıkları zamanların belirlenmesi amaçlanmaktadır. Bu sayede işletme kendileri için kritik olan hedef kitleyi daha iyi tanıyıp, her gruba ya da kişiye özel uygulamalar sunarak müşteri memnuniyetini artırabilecektir.

Eğitim danışmanlarının, müşteriler bilgi amaçlı geldiğinde veya telefonla bilgi amaçlı aradıklarında tuttuğu veriler bir araya getirilerek çalışma için gerekli veriler bir araya getirilmiştir. Henüz sektörde yeni olan şirketin ele alınmasının sebebi rakip şirketlere göre sektörde yeni bir danışmalık şirketi olmasından kaynaklı hızlı bir sekilde ilerlemeyi amaçlamaktadır.

Çalışma için 727 müşterinin verileri alınmış, 8 tane bağımsız değişken, 1 tanede bağımlı değişken veri setine dâhil edilmiştir. Bağımlı değişken şirkete kayıt olup olmaması ile ilgiliyken, bağımsız değişken kayıt olanların ya da sadece danışmanlık alanların bilgileri alınmıştır.

Çalışmada yer alan bağımsız değişkenler ve kategorileri aşağıda tanımlanmıştır.

Bağımsız değişkenlerde 12 yaş altı müşterilerde karar verici aileler olarak tanımlanmıştır.

İlk bağımsız değişken cinsiyet üzerinden yapılmıştır. Numerik olarak değişkenin değerleri:1-erkek/ 2- kadın olarak ayrılmıştır.

2. bağımsız değişken de 12-32 yaş aralığında olan kişilerin yaş aralıklarına göre ayrımı yapılmıştır. Numerik olarak değişkenin değerleri: 1- (12-18) yaş aralığı / 2-(18-22) yaş

aralığı / 3-(22-24) yaş aralığı / 4-(24-28) yaş aralığı/ 5-(28-32) yaş aralığı olarak ayrılmıştır.

3. bağımsız değişken bireyin güncel durumu ile ilgilidir. Bu değişken numerik olarak değeri:1-öğrenci/ 2- çalışan olarak ayrılmıştır.

4. bağımsız değişken bireyin ilgilendiği alan başka bir tanımla bölümü olarak ayrılmıştır. Numerik olarak değişkenin değerleri:1-mühendislik / 2-sağlık / 3-işletme / 4- uluslararası ilişkiler / 5- uluslararası ticaret ve lojistik/ 6- İngiliz dili ve edebiyat/ 7- ortaokul ve lise bölümü olarak ayrılmıştır.

5. bağımsız değişken danışmanlık firması ile iletişime geçen bireylerin hangi şehirlerden iletişime geçtiği ile ilgili olarak ele alınmıştır. Numerik olarak değişkenin değerleri: 1- Ankara/ 2- İstanbul/ 3- Karabük / 4- İzmir / 5- Kocaeli/ 6- Sakarya olarak ayrılmıştır.

6. bağımsız değişken bireyin daha önce yurt dışı deneyimi olup olmamasına göre ayrılmıştır. Numerik olarak değişkenin değerleri: 1-var / 2- yok olarak ayrılmıştır.

7.bağımsız değişken bireyin danışmanlık firması ile hangi paket program için iletişime geçtiği ile ilgilidir. Numerik olarak değişkenin değerleri: 1- Dil okulu, 2- Master Eğitimi 3- Summer Work and Travel/ 4- Vize Danışmanlığı/ 5 – Yaz Kampı olarak ayrıştırılmıştır.

8. ve son bağımsız değişken bireyin hangi tarih aralığında danışmanlık firması ile iletişime geçtiği konusunda ayrılmıştır. Numerik olarak değişkenin değerleri: 1- Ocak/ 2- Şubat/ 3- Mart/ 4- Nisan/ 5 –Mayıs/ 6- Haziran/ 7- Temmuz/ 8- Ağustos/ 9- Eylül/ 10- Ekim/ 11-Kasım/ 12- Aralık olarak ayrılmıştır.

9. olarak bağımlı değişken için 1 sene sonunda bireylerin kayıt olma ve olmama durumu hedef değişken olarak belirlenmiştir. Numerik olarak değişkenin değerleri: 1- Kayıt olmadı/ 2- Kayıt oldu olarak numaralandırılmıştır.

Tablo 4.1. Örneklemin Demografik Özellikleri.

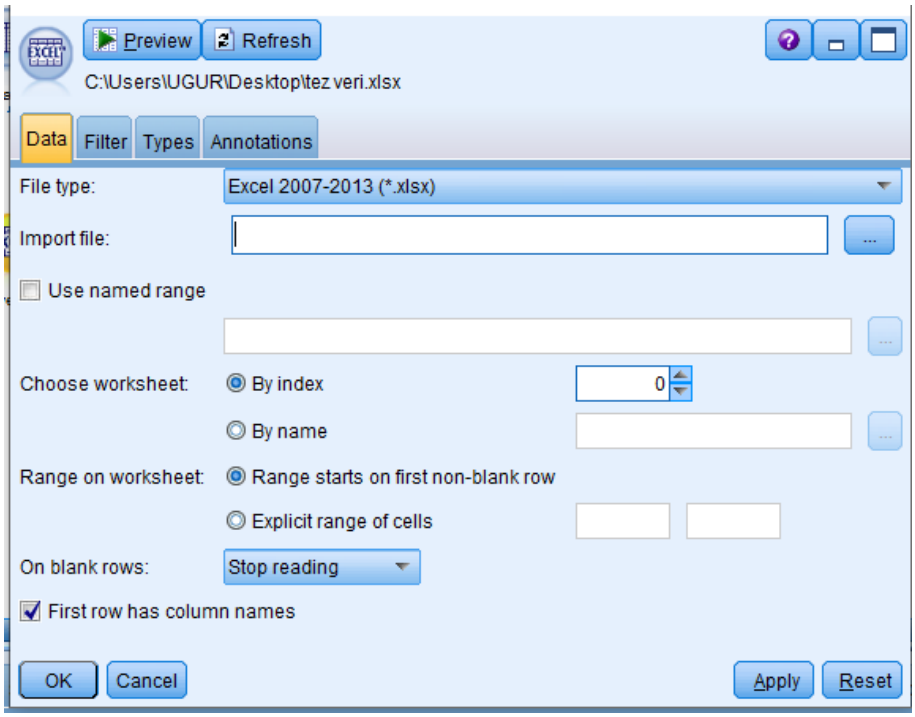
Değişken	Belirleyici Faktör	Sayı	Yüzde
Cinsiyet	Erkek	339	0,466
	Kadın	388	0,534
Yaş Kategorisi	12-18	95	0,131
	18-22	287	0,395
	22-24	151	0,208
	24-28	98	0,135
	28-32	96	0,132
Alan	Mühendislik	77	0,106
	Sağlık	75	0,103
	İşletme	119	0,164
	Uluslararası İlişkiler	231	0,318
	Uluslararası Ticaret ve Lojistik/	32	0,044
	İngiliz Dili ve Edebiyatı	95	0,131
Statü	Ortaokul-Lise	98	0,135
	Çalışan	263	0,362
	Öğrenci	464	0,638
Şehir	Ankara	84	0,116
	İstanbul	129	0,177
	Karabük	136	0,187
	İzmir	123	0,169
	Kocaeli	143	0,197
	Sakarya	112	0,154
Yurt Dışı Deneyimi	Var	335	0,461
	Yok	392	0,539
Başvurduğu Program	Yurt Dışı Dil Okulu	233	0,32
	Yüksek Lisans	110	0,151
	Work and Travel	177	0,243
	Vize Danışmanlığı	150	0,206
	Yaz Kampı	57	0,078
Aranılan Tarih	Ocak	43	0,059
	Şubat	50	0,069
	Mart	81	0,111
	Nisan	57	0,078
	Mayıs	93	0,128
	Haziran	67	0,092
	Temmuz	45	0,062
	Ağustos	102	0,14

Tablo 4.1. (Devamı) Örneklemin Demografik Özellikleri.

Eylül	48	0,066
Ekim	57	0,078
Kasım	63	0,087
Aralık	21	0,029

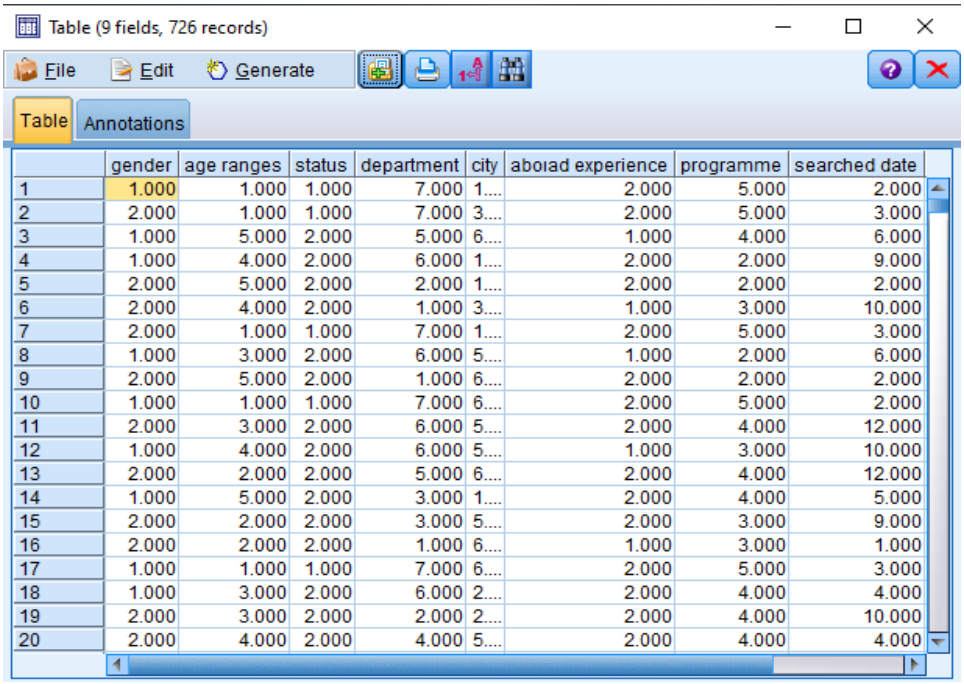
4.2. Spss Modeler (Clementine) ile Verilerin Hazırlanması

Uygulama adımında, bütün olarak görsel bir modelleme imkânı sunan SPSS Clementine programı kullanılmıştır. SPSS Clementine programı, veri madenciliği çözümlerinde hem istatistik kökenine dayanan algoritmaları hem de yapay zekâ kökenine dayanan algoritmaları görsel bir şekilde sunmaktadır [57].



Şekil 4.1. Verinin Yüklenmesi.

Çalışmada kullanılan veri seti Exceldir. Sources modülünden Excel kaynağı seçilir ve Import file'dan dosya kaynağını vermek için veriler tanımlanır. Verilerin kontrolü amacı ile preview modülü kullanılır ve kayıtlar kontrol edilir. Veri setimiz excel olduğu için akış diyagramına excel fonksiyonu eklenerek başlanır. Veri setinin yüklenmiş halini örneği Şekil 4.2 de gösterilmiştir.



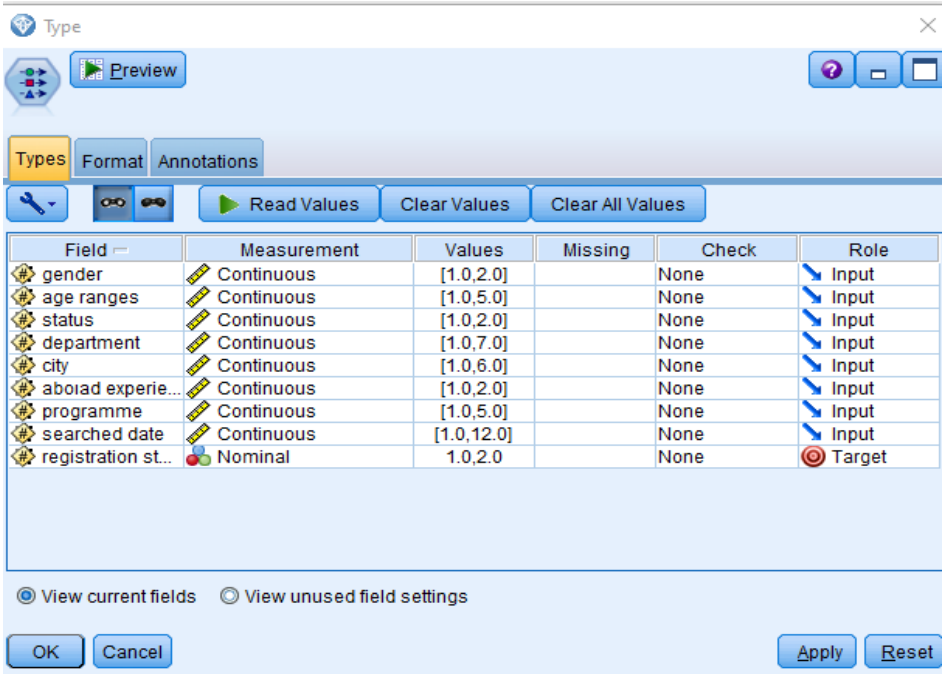
	gender	age ranges	status	department	city	aboıad experience	programme	searched date
1	1.000	1.000	1.000	7.000	1....	2.000	5.000	2.000
2	2.000	1.000	1.000	7.000	3....	2.000	5.000	3.000
3	1.000	5.000	2.000	5.000	6....	1.000	4.000	6.000
4	1.000	4.000	2.000	6.000	1....	2.000	2.000	9.000
5	2.000	5.000	2.000	2.000	1....	2.000	2.000	2.000
6	2.000	4.000	2.000	1.000	3....	1.000	3.000	10.000
7	2.000	1.000	1.000	7.000	1....	2.000	5.000	3.000
8	1.000	3.000	2.000	6.000	5....	1.000	2.000	6.000
9	2.000	5.000	2.000	1.000	6....	2.000	2.000	2.000
10	1.000	1.000	1.000	7.000	6....	2.000	5.000	2.000
11	2.000	3.000	2.000	6.000	5....	2.000	4.000	12.000
12	1.000	4.000	2.000	6.000	5....	1.000	3.000	10.000
13	2.000	2.000	2.000	5.000	6....	2.000	4.000	12.000
14	1.000	5.000	2.000	3.000	1....	2.000	4.000	5.000
15	2.000	2.000	2.000	3.000	5....	2.000	3.000	9.000
16	2.000	2.000	2.000	1.000	6....	1.000	3.000	1.000
17	1.000	1.000	1.000	7.000	6....	2.000	5.000	3.000
18	1.000	3.000	2.000	6.000	2....	2.000	4.000	4.000
19	2.000	3.000	2.000	2.000	2....	2.000	4.000	10.000
20	2.000	4.000	2.000	4.000	5....	2.000	4.000	4.000

Şekil 4.2. Veri Seti.

Bu kısımda sadece ham veriler dâhil edilir ve isteğe bağlı olarak tip belirleme işlemi yapılır. Bu çalışmada bu bölümde herrhangi bir filtreleme veya tip belirleme işlemi yapılmamıştır. Table fonksiyonu burada veri setini gözden geçirmek için kullanılmıştır. Bu kısımda verinin okunmasını ve herhangi bir kayıp veri içerip içermediğini kontrol etmek hedeflenir.

4.2.1. Type modulü

Field Operation bölümünde yer alan Type modulü ile aktarılan veriler arasında bağlantı kurulur. Connect seçeneği ile akış diyagramının elemanları birbirine bağlanır. Read values yardımı ile veride ne olup olmadığı okunur ve değişkenlerin hangi aralıkta hangi değerleri aldığı görülmüş olup, bu değerler ordinal ölçekli, nominal ölçekli değerler olabilmektedir. Problemin tanımına göre değişkenlerin kategorisi belirlenir. Bu kategoriyi belirlemek için Type modulün de sonunda yer alan Role bölümünde; bir bireyin kayıt olup olmama durumunu sistemin target verisi olması sebebiyle "kayıt durumu" out verisi, geriye kalanlar ise bu süreçte input bir diğer ismi ile girdi verisi olarak programa dâhil edilir. Tüm bu yapılardan sonra program girilen verilerin ne olduğunu ve nasıl kullanılacağını ve hangi amaçta olduğunu bilir durumdadır. Type modülü örneği şekil 4.3 de gösterilmiştir.

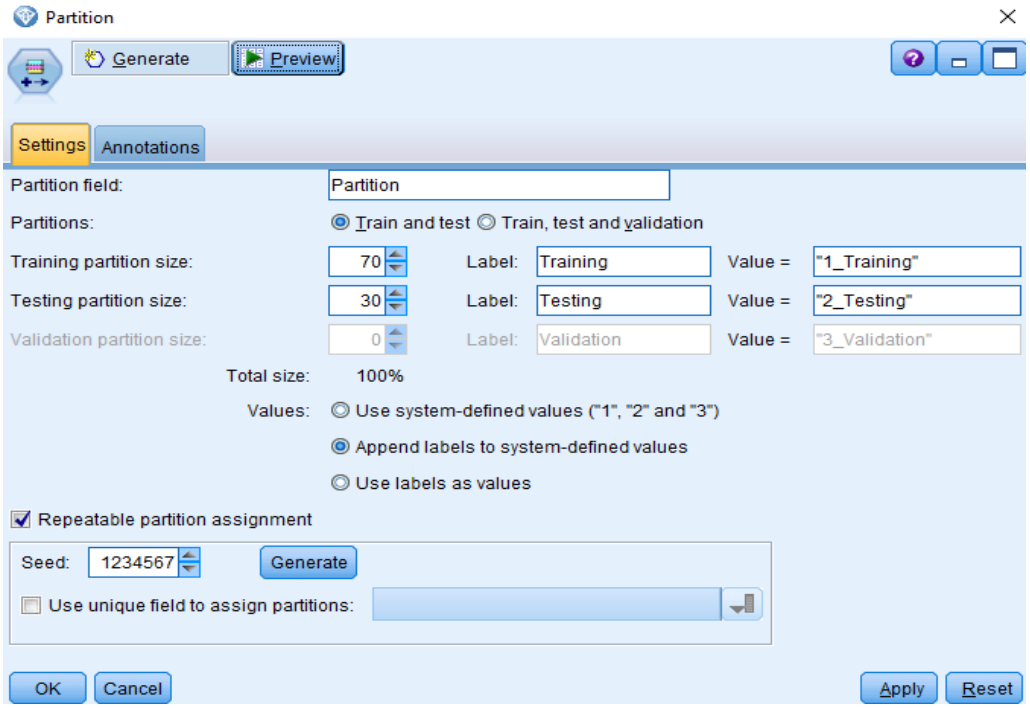


Şekil 4.3. Type Modülü Örneği.

Bu işlem de özet olarak 727 bireyin verilerini kullanarak bir sınıflandırma modeli kurarak yeni gelecek kişinin bu değişkenlere bakarak kayıt olup olmayacağının tespitini ortaya koyacak bir sınıflandırma modeli kurulması amaçlanmaktadır.

4.2.2. Partition modülü

Model oluşturulurken algoritmaya eğitim ve test verisi veya doğrulama verisi olmak üzere 3 grup veri verilmesi gerekmektedir. Eğitim verileri yardımıyla değişkenler arası verileri öğrenip, test kümesi yardımıyla da ilişkinin performansını değerlendirilir. Bu durumun sınıflandırma modeli için ayrılması gerekmektedir



Şekil 4.4. Partition Modülü Örneği.

Partition modülü ile eğitim ve test kümesinin yüzde kaçlık dilimini ayırdığımız belirlenir. Çalışmada kullanılan veriler; çalışmada kullanılan her iki metot için de %75 training %25 testing olarak ayrılmıştır. Literatürde en fazla tercih edilen model % 70 ve %30 ayrımlardır. Bu çalışmada da literatürde en fazla tercih edilen yapıya yakın değerler ele alınmıştır. En iyi modeli kurmak için bu değerler değiştirilebilir. Eğitim verisi ayırmak, test verisi ayırmak, kullanacağımız sınıflandırma algoritmasına karar vermek, değişken sayısını belirlemek, veri büyüklüğüne karar vermek vb. kriterler veri madenciliğinde karar verilmesi gereken belirleyicilerdir. Bu adımdan sonra problem çeşidine, değişkenlere göre algoritma seçimi yapılır.

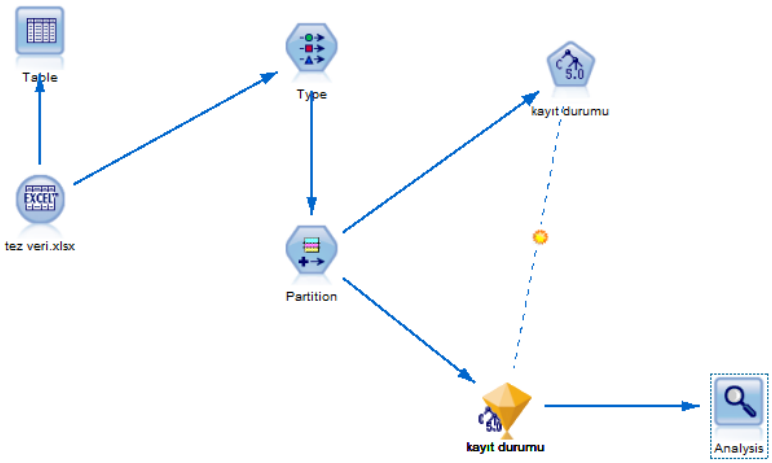
Modeling fonksiyonun içinde kullanacağımız karar ağaçları algoritmaları yer almaktadır. Algoritmalar belirlenirken ne tür bir sınıflandırma ne karar ağacı ortaya koymak istediğimize göre algoritma seçilir. Çalışma da veri madenciliği karar ağacı algoritmalarından C.5.0 ve C&R algoritmaları ile kurulan karar ağacı modellerine yer verilmiştir

4.3. C5.0 Algoritması

Bu bölümde yurt dışı eğitim danışmanlık firmasını arayıp bilgi alan ve verileri tutulan bireylerin daha sonra kayıt olma ve olmama durumuna göre veriler ayrıştırılmıştır.

Kayıt durumu hedef değişken olarak belirlenmiş ve veri madenciliği karar ağacı algoritmalarından C.5.0 ve C&R algoritmaları ile kurulan karar ağacı modellerine yer verilmiştir.

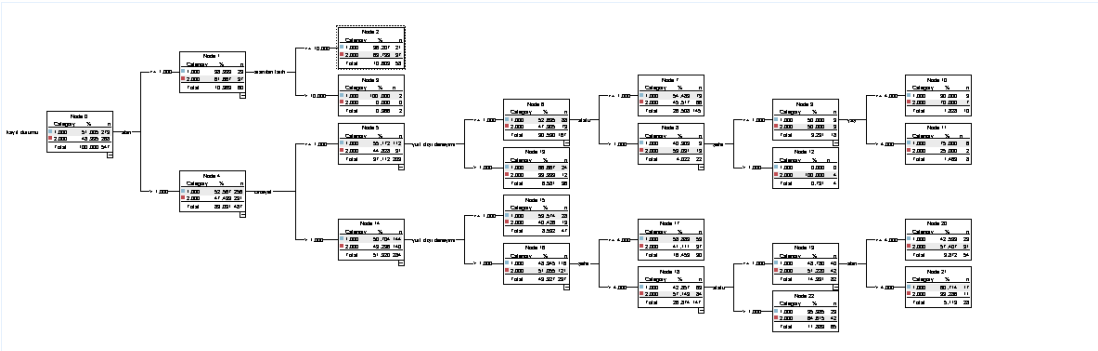
Çalışma Kapsamında yurt dışı eğitim danışmanlık firması ile iletişime geçen, firmaya kayıt olsun veya olmasın 727 kişinin verileri kullanılmıştır. Yanıt değişkeni "kayıt durumu" olarak belirlenmiştir. Kayıt durumu belirlenirken 8 kategorik değişken kullanılmıştır. Model için IBM SPSS Modeler programında yapılan işlemler bölüm 4.2 de her bir modül anlatılmış olup C5.0 Modelleme süreci şekil 4.5 de gösterilmiştir.



Şekil 4.5. IBM SPSS Modeler C5.0 Modelleme Süreci.

4.3.1. C5.0 algoritması karar ağacı

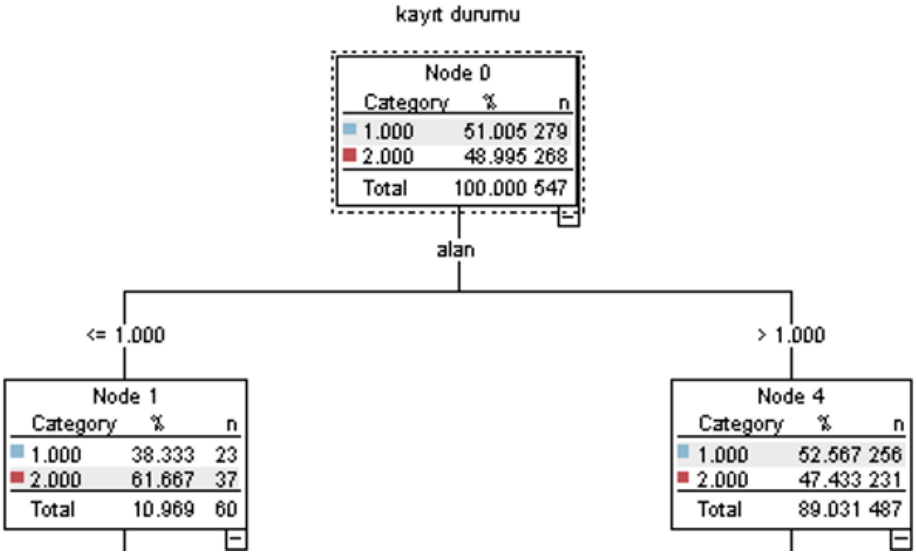
Ağacın ilk kısmında 727 kayıttan %75 training % 25 testing olarak ayrılmıştır. Entropi tabanlı C5.0 algoritmasının karar ağacı Şekil 4.6'da gösterilmişrtir.



Şekil 4.6. C5.0 Algoritması Karar Ağacı.

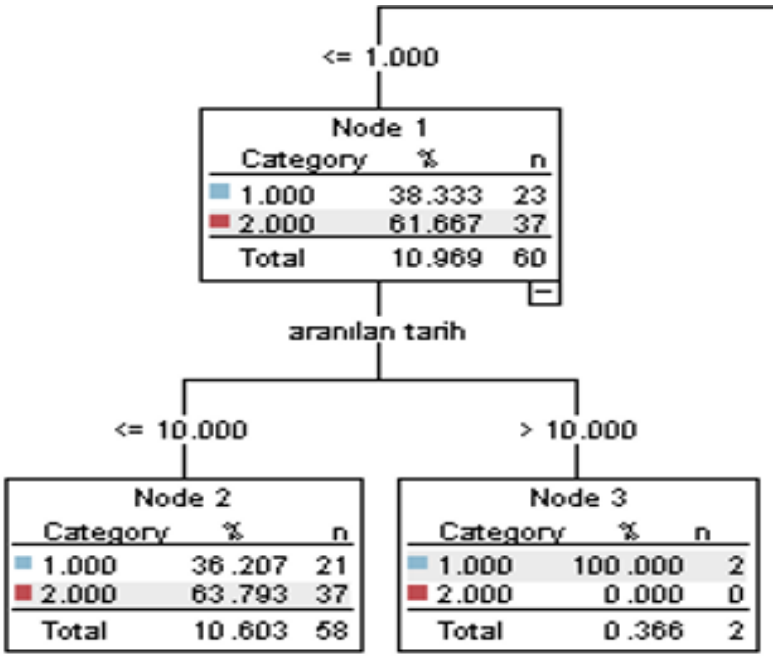
Ağacın büyüklüğünden dolayı sadece ağacın kapsamlı hali ve dikey hali çok fazla yer kaplaması sebebiyle yan haline yer verilmiştir.

Karar ağacından çıkan kurallara geçmeden önce, modelleme sonucunda meydana gelen karar ağacına ait bir takım yorumlamalara yer verilmiştir.



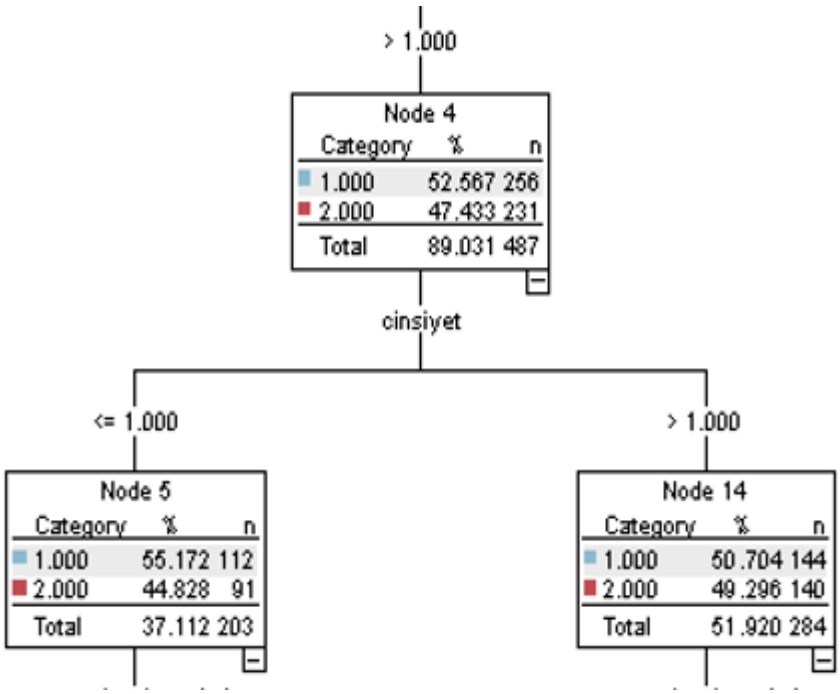
Şekil 4.7. Karar Ağacı Dallanması Örneği 1.

C5.0 karar ağacı modelinde ağaç ilk ayrımını alan değişkenine göre yapmıştır. Ayrımı yaparken 727 kişinin %75'i 547 kişiyi değerlendirmeye almıştır. Bu ayrımı da alanı belirlenen numerik değişkenlerden 1'e eşit olanlar ve 1'den büyük olanlar olarak ayırmıştır. Training kısmında yer alan 547 bireyden 60 'ı alan 1 eşit olanlar yani alanı mühendislik olanlar ve geriye kalan 487 kişi alanı diğer bölümler olarak ayrılmıştır.



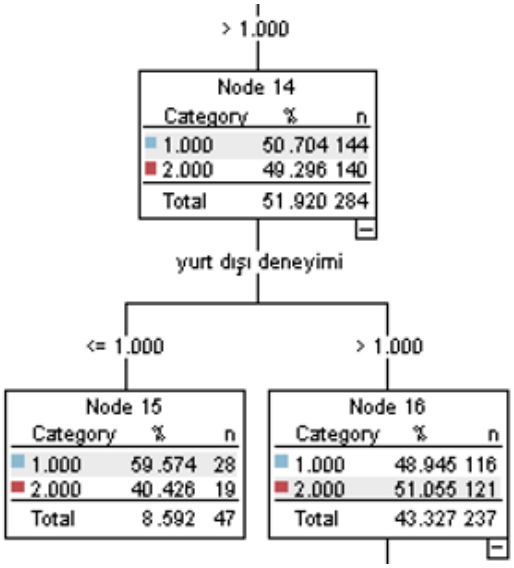
Şekil 4.8. Karar Ağacı Dallanması Örneği 2.

Alanı 1 numerik değişkenine sahip olan 60 kişiyi kendi arasında aradıkları tarihe göre ayrım yapmıştır. Aranılan tarih ayrımını aradıkları tarih 10 numerik değişkenine sahip olan yani ekim ayından önce ve sonra ayıranlar olarak devam etmiştir.



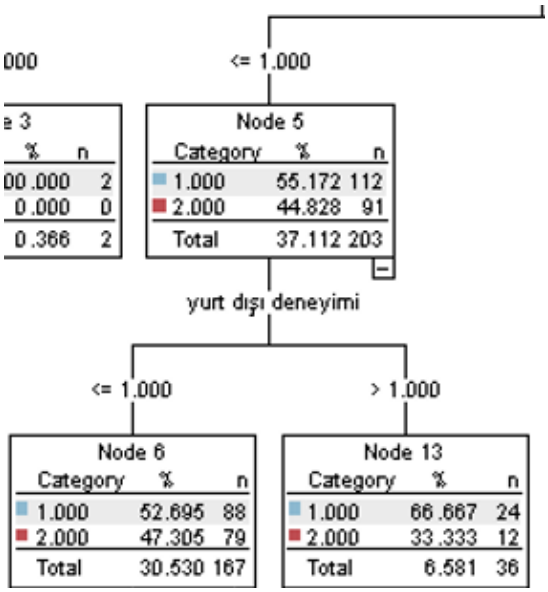
Şekil 4.9. Karar Ağacı Dallanması Örneği 3.

Alanı 1 den büyük olan 487 kişiyi kendi içerisinde cinsiyeti kadın olan 284 kişi ve erkek olan 203 kişi olarak 2 gruba ayırmıştır.



Şekil 4.10. Karar Ağacı Dallanması Örneği 4.

Alanı 1 den büyük olan kişileri cinsiyeti kadın ve erkek olacak şekilde ayırdıktan sonra kadın ve erkekleri kendi içerisinden yurt dışı deneyimi olup olmama durumuna göre ayırmıştır. Cinsiyeti kadın olan 284 kişiden 47 kişisinin yurt dışı deneyimin olduğu geriye kalan 237 kişiden yurt dışı deneyimi olmadığı görülmüştür.

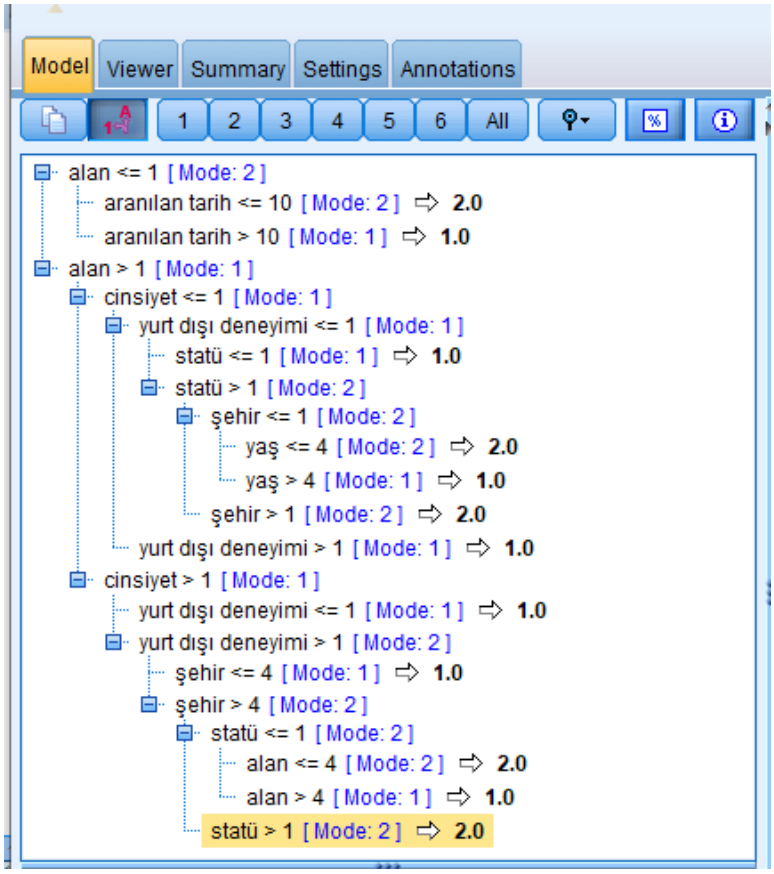


Şekil 4.11. Karar Ağacı Dallanması Örneği 5.

Aynı şekilde cinsiyeti 1 numerik değişkenine yani erkek olan 203 kişiden 167 kişisinin yurt dışı deneyiminin olduğu 36 kişisinin yurt dışı deneyimi olmadığı sonucuna varılmuştur.

4.3.2. C5.0 algoritması kuralları

Program çalıştırıldığında oluşturulan karar ağacının oluşturulma kuralları aşağıdaki karar modeline göre oluşmaktadır. Öncelikle ilk ayrımı oluşturacağımız sınıflandırma kuralını alan değişkeni üzerinden yapıyor. Program çıktısı öncelikle 1. Level karar ağacını gösteriyor. Eğer 3 düzey ilerlemek istenirse 3 butonunu kullanrak bütün karar modelini görmek için de All sekmesine tıklayarak görülmektedir.

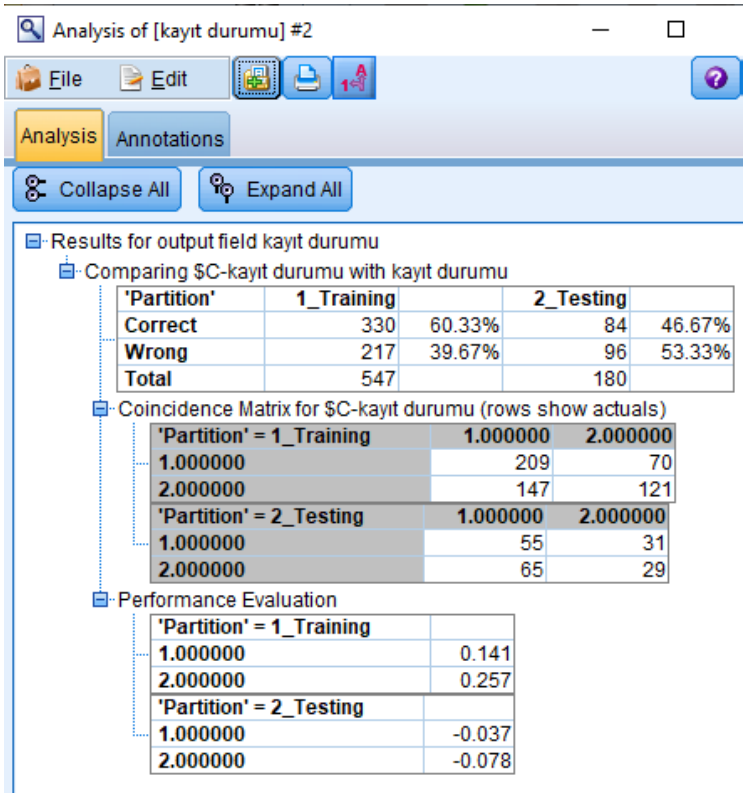


Şekil 4.12. C5.0 Algoritması Kuralları.

İlk kurala göre alanı mühendislik olan ve 10. Aydan önce arayanlardan %63 ü kayıt oldu olarak sınıflandırılmıştır. Alanı mühendislik dışında olan, cinsiyeti erkek olan, yurt dışı deneyimi olmayan olan, Ankara şehrinden arayan 24 yaşından küçük öğrencilerin %70'i kayıt oldu olarak sınıflandırılmıştır. Alanı mühendislik dışında olan, cinsiyeti erkek olan, yurt dışı deneyimi olmayan olan ve İstanbul şehrinden

arayan öğrencilerin hepsi kayıt oldu olarak sınıflandırılmıştır. Alanı mühendislik dışında olan, erkek bireylerin yurt dışı deneyimi olmayan Sakarya şehrinden arayan ve çalışan grubu % 64'ü kayıt oldu olarak sınıflandırılmıştır. Alanı 1'den büyük 4'e küçük eşit olan yani numerik değişken karşılığı 2 ve 3 değerinde sağlık ve işletme olan, yurt dışına deneyimine sahip Sakrya şehrinden arayan öğrencilerin % 57'si kayıt oldu olarak sınıflandırılmıştır.

4.3.3. C5.0 algoritması performans değerlendirme



'Partition'	1_Training		2_Testing	
Correct	330	60.33%	84	46.67%
Wrong	217	39.67%	96	53.33%
Total	547		180	

'Partition' = 1_Training	1.000000	2.000000
1.000000	209	70
2.000000	147	121
'Partition' = 2_Testing	**1.000000**	**2.000000**
1.000000	55	31
2.000000	65	29

'Partition' = 1_Training	
1.000000	0.141
2.000000	0.257
'Partition' = 2_Testing	
1.000000	-0.037
2.000000	-0.078

Şekil 4.13. C5.0 Algoritması Performans Değerlendirmesi.

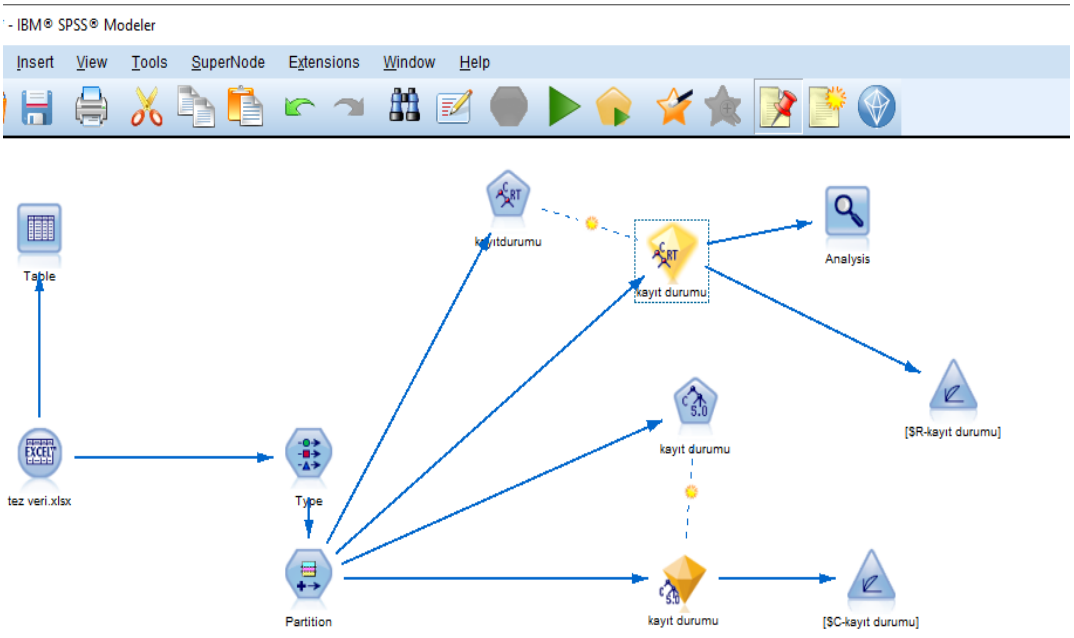
Modelin performans değerini training bölümü için yorumlandığında, 547 kişinin bulgularının bu modele göre tanımlamış olduğu varsayılırsa 330 kişiyi doğru olarak model bulabilecektir. Eğer kişi kayıt oldu ise kayıt olmuş olarak, kayıt olmadıysa kayıt olmamış olarak doğru bir şekilde bulabilecektir. Bu durum kurulan modelin %60.33'lük bir başarısı olduğunu göstermektedir. Sınıflandırma modelinin daha detaylarında ise kayıt olmayan 209 kişiyi kayıt olmadı şeklinde doğru yorumlarken, kayıt olmayan 70 kişiyi kayıt oldu olarak yanlış yorumlamıştır. Kayıt olan 121 kişiyi

kayıt oldu olarak doğru sınıflandırırken, kayıt olan 147 kişiyi kayıt olmadı olarak yanlış yorumlamıştır.

Modelin performans değerinin test verisi için yorumlandığında kayıt olmayan 55 kişiyi kayıt olmadı olarak doğru yorumlarken, kayıt olmayan 31 kişiyi kayıt oldu olarak yanlış yorumlamıştır. Kayıt olmayan 29 kişiyi kayıt olmadı olarak kayıt olmadı olarak doğru sınıflandırırken, kayıt olmayan 65 kişiyi kayıt oldu olarak doğru sınıflandırmıştır. Bu durumda model training kısmı bölümü için %69,74 başarı elde ederken, test verisi için %46,67 oranında bir başarı elde etmiştir.

4.4. C& R Algoritması

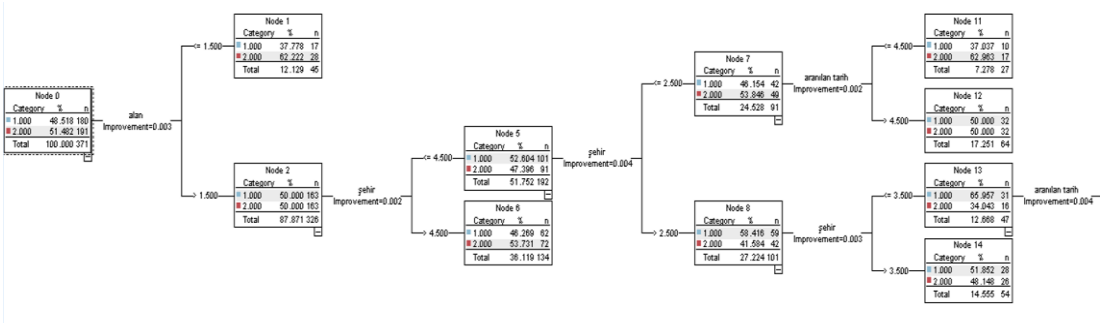
Modeller için IBM SPSS Modeler programında yapılan işlemler Şekil 4.14'de gösterilmiştir. Bu bölümde C&R algoritması ile elde edilen sonuçlara yer verilecektir. Bu arayüz de birden fazla algoritma seçimi aynı ekranda gösterilmiştir.



Şekil 4.14. IBM SPSS Modeler C5.0 & C&R Modelleme Süreci.

4.4.1. C& R algoritması ağaç yapısı

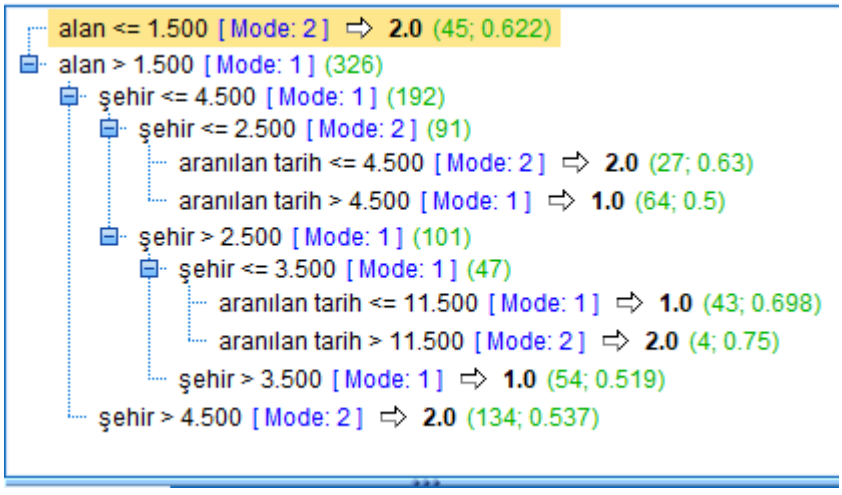
Ağacın ilk kısmında 727 kayıttan %75 training % 25 testing olarak ayrılmıştır. Entropi Ağacın büyüklüğünden dolayı sadece ağacın kapsamlı hali ve dikey hali çok fazla yer kaplaması sebebiyle yan haline yer verilmiştir.



Şekil 4.15. C&R Algoritması Ağaç Yapısı.

Ağacın ilk kısmında 727 kayıttan %75 training % 25 testing olarak ayrılmıştır. Entropi Ağacın büyüklüğünden dolayı sadece ağacın kapsamlı hali ve dikey hali çok fazla yer kaplaması sebebiyle yan haline yer verilmiştir.

4.4.2. C&R algoritması kuralları



Şekil 4.16. C&R Algoritması Kuralları.

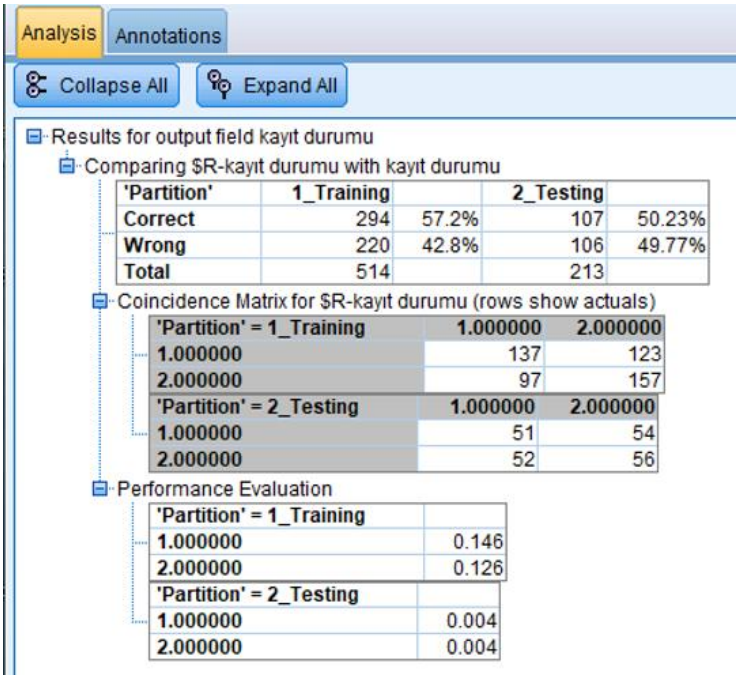
C&R algoritması ağacı bölmeye alan değişkeninden başlamıştır. İlk ayrım alanı mühendislik olan ve diğer bölümler olan (Sağlık, İşletme, Uluslararası İlişkiler, Uluslararası ticaret ve lojistik, İngiliz Dili ve Edebiyatı, Ortaokul Lise) olarak

ayırmıştır. Alanı mühendislik olan kişilerden %62 si kayıt oldu olarak sınıflandırılmıştır. Alanı mühendislik dışında olan ve Sakarya şehrinden katılım sağlayan kişilerin %53'ü kayıt oldu olarak sınıflandırılan gruba girmiştir.

Alanı Sağlık, İşletme, Uluslararası İlişkiler, Uluslararası ticaret ve lojistik, İngiliz Dili ve Edebiyatı, Ortaokul Lise olan, 1 ve 2 şehir numerik değişkenlerinin karşılığı olan Ankara İstanbul şehirlerinden katılım sağlayan ve ilk 4 ayda arayan arayan kişilerin kişilerden %63'ü kayıt oldu olarak sınıflandırılırken, 4.Aydan sonra arayanların %50'si kayıt olmadı olarak sınıflandırılmıştır.

Alanı Sağlık, İşletme, Uluslararası İlişkiler, Uluslararası ticaret ve lojistik, İngiliz Dili ve Edebiyatı, Ortaokul Lise olan, 3 numerik şehir değişkeninin karşılığı İzmir olan ve 11.Ay dahil olmak üzere öncesinde arayan kişilerin %69'u kayıt olmadı olarak sınıflandırılırken, 12. Ayda arayan kişilerin %75'i kayıt oldu olarak sınıflandırılmıştır.

4.4.3. C&R algoritması performans değerlendirme



'Partition'	1_Training		2_Testing	
Correct	294	57.2%	107	50.23%
Wrong	220	42.8%	106	49.77%
Total	514		213	

'Partition' = 1_Training	1.000000	2.000000
1.000000	137	123
2.000000	97	157
'Partition' = 2_Testing	**1.000000**	**2.000000**
1.000000	51	54
2.000000	52	56

'Partition' = 1_Training	
1.000000	0.146
2.000000	0.126
'Partition' = 2_Testing	
1.000000	0.004
2.000000	0.004

Şekil 4.17. C5.0 Algoritması Performans Değerlendirmesi.

Veri madenciliği karar ağacı algoritmalarında modelin performans değeri farklı yöntemler ile değerlendirilebilir. Bu çalışmada literatürde çok fazla rastlanmayan SPSS Modeler programında analysis modülü performans değerlendirme yapılmıştır.

Analysis modülü ile kurulan modelin trainig ve testing değerlendirmelerinde modelin doğru ve yanlış tahminlemelerinin oranını göstermektedir.

Modelin performans değerinde detaylı bir coincidence matrix elde edilir. Performans değerini training bölümü için yorumlandığında, 514 kişinin bulgularının bu modele göre tanımlamış olduğu varsayılırsa 294 kişiyi doğru olarak model bulabilecektir. Eğer kişi kayıt oldu ise kayıt olmuş olarak, kayıt olmadıysa kayıt olmamış olarak doğru bir şekilde bulabilecektir. Bu durum kurulan modelin %57.22'lik bir başarısı olduğunu göstermektedir. Testing bölümünde yer alan bireylerin bulguları ortaya çıkan modele göre yorumlandığında %50 oranla doğru bulabilecektir.

Sınıflandırma modelinin daha detaylarında ise training bölümündeki bireylerden, kayıt olmayan 137 kişiyi kayıt olmadı şeklinde doğru yorumlarken, kayıt olmayan 123 kişiyi kayıt oldu olarak yanlış yorumlamıştır. Kayıt olan 157 kişiyi kayıt oldu olarak doğru sınıflandırırken, kayıt olan 97 kişiyi kayıt olmadı olarak yanlış yorumlamıştır.

Modelin performans değerinin test verisi için yorumlandığında kayıt olmayan 51 kişiyi kayıt olmadı olarak doğru yorumlarken, kayıt olmayan 57 kişiyi kayıt oldu olarak yanlış yorumlamıştır. Kayıt olmayan 56 kişiyi kayıt olmadı olarak kayıt olmadı olarak doğru sınıflandırırken, kayıt olmayan 52 kişiyi kayıt oldu olarak doğru sınıflandırmıştır. Bu durumda model training kısmı bölümü için %57.22 başarı elde ederken, test verisi için %50,23 oranında bir başarı elde edilmiştir.

5. SONUÇ

Ülkemizde özellikle son yıllarda yurt dışına seyahat talebi önemli ölçüde artmıştır. Bireyler yurt dışına birden fazla amaçla seyahat etmektedir, özellikle dil öğrenmek amacıyla yurt dışına gidenlerin sayısı her geçen gün artmaktadır. Bu artış durumu da yurt dışı eğitim-danışmanlık kurumlarının Türkiye'de öneminin artmasına sebebiyet vermiştir. Aynı zamanda bireyler en kaliteli ve en profesyonel hizmeti almayı hedeflemektedir. Bu durum yurt dışı eğitim danışmanlığı kurumları arasında da rekabete yol açmaktadır. Bu çalışmada yer alan kurum diğer kurumlara göre yeni kurulmuş olduğu için işletmenin müşteri portföyünün belirlenmesi ve buna göre iyileştirme çalışmalarının yapılması, yeni kurulmuş bir firma olduğu için pazardan pay almak ve fazla harcama yapmamak için daha doğru kararlar vermesi hedeflenmektedir.

Yurt dışı eğitim firmasında yapılan çalışmada 8 adet bağımsız değişken ve 1 adet bağımlı değişken kullanılarak müşteri kayıt durumunu etkileyen faktörler belirlenerek işletme kaynakları ile işletmenin tekliflerine en çok yanıt verecek tüketici gruplarını tespit edilmesi amaçlamıştır. Çalışmada SPSS Modeler programı kullanılmış olup karar ağacı yöntemlerinden C5.0 ve C&R algoritmaları kullanılmıştır. C5.0 algoirtması ağaç ayrımı yaparken entropi değerleri belirlemektedir ve entropi değerlerine göre ayrım yapmaktadır. İlk olarak C5.0 algoritmasının modelinden ilk ayrım alan değişkenine göre yapılmıştır ve bunun devamında aranılan tarih, cinsiyet ve yurt dışı deneyimi kriterleri yer almıştır. Bu durumda model çıktı sonuçlarına göre kayıt olma durumunda etkili olan en önemli faktörlerin; alan, aranılan tarih ve yurt dışı deneyimi olma durumu yer almıştır. Özellikle sınflandırma gruplarından %70'i ve %64'ü kayıt oldu olarak ayrılan kümenin her ikisinin de ortak noktası bireylerin alanının mühendislik dışında olması, erkek birey olması ve yurt dışı deneyiminin olmaması, belirleyici faktörklerden olmuştur. Özellikle alanı mühendislik dışında olan, cinsiyeti erkek olan, yurt dışı deneyimi olmayan olan ve İstanbul şehrinden arayan öğrencilerin de hepsinin danışmanlık şirketine kayıt olması C5.0 algoritması için yine belirleyici bir müşteri profili olma ihtimalini taşımıştır. Bu durumda danışmanlık şirketinin hedef kitlesi alanı mühendislik dışında olan, İstanbul şehrinde yaşayan ve daha önce yurt dışı deneyimi olmayan bireyler olarak belirlenebilir. Bu

bilgiler dahilinde danışmanlık firmasının iyileştirme çalışması olarak ilerleyen zamanlarda fazla rağbet gören şehirlere gerekirse ek şube açması, yurt dışı deneyimi olmayan kişilere şirket tanıtımı ve yer alan programları için gerekli çalışmaları yapması ve öğrenciler için özellikle üniversitelerin öğrenci kulüpleri ile entegre bir şekilde çalışması ve bu sayede firma tanıtımı için alınabilecek aksiyonlardan biri olabilir.

C&R algoritması modelleme sonucuna göre kayıt olup olmama durumunu etkileyen en önemli faktörler alan, şehir ve aranılan tarih olarak ayrılmıştır. C&R algoritması da ilk ayrıma C5.0 da olduğu gibi alanı mühendislik ve mühendislik dışında olanlar olarak ayrılmıştır. Alanı mühendislik dışında olanları şehir değişkenine göre kendi içeridinde ayrılmıştır. Şehir değişkeninin ilk ayrımın Kocaeli, Sakarya şehrinden arayan ve diğerrleri olarak ayırmıştır. Geri kalanları da kendi içerisinde Ankara, İstanbul ve Karabük, İzmir olarak ayırmıştır. Bu ayrımdan sonra her iki durum içinde aranılan tarih ayırma kriteri olmuştur. Bu kurallardan sonra oluşan sınıflandırmalarda alanı mühendislik olan kişlerin %62 si kayıt oldu olarak sınıflandırılırken ; alanı mühendislik dışında olan Ankara, İstanbuldan katılım sağlayan ilk 4 ayda arayan kişilerin %63 ü kayıt oldu olarak sınıflandırılarak ve alanı mühendislik dışında olan, Kocaeli-Sakarya bölgesinden katılım sağlayan kişilerin %53 ü kayıt oldu olarak sınıflandırılarak alan ve şehir değişkenleri en belirleyici kritlerler arasında yer almıştır ve bunun devamında belirleyici faktör aranılan tarih olarak devam etmiştir.

İki algoritmanın sonucu karşılaştırıldığında her iki algoritma içinde belirleyici faktörler alan, şehir ve aranılan tarih olması olması dikkat çekmiştir.

Ele alınan çalışma da bireyin kayıt olup olmama durumuna odaklanılmıştır. Çalışmanın yapılma amaçları daha önce literatürde bu tarz bir çalışmaya rastlanmaması, kullanılan teknolojinin globalleşmesi ile günlük yaşantımızda dil kullanımını giderek artması, buna bağlı olarak yurt dışı eğitim kurumlarına rağbetin artması çalışmanın yapılmasın da etkin rol almıştır. Bu durumda hem hizmet sektörü için hem de müşteriler için karşılıklı bir fayda olabileceği düşünülmektedir. Çalışmayı geliştirmek için daha fazla bireyin verilerinin olması, farklı karar ağacı algoritmalarının ve ya farklı veri madenciliği metotlarının kullanılması, algoritmanın sonucunu etkileyen değişkenlerin değiştirilmesi ile sonuçlar karşılaştırılıp farklı

sonuçlar elde edilip yorumlanabilir. Benzer şekilde kayıt durumunu etkileyebilecek farklı kriter de ele alınıp analiz edilebilir.

KAYNAKLAR

[1] YÖK. (2020a). Yükseköğretim Kurulu. https://COVID19.yok.gov.tr/Documents/alinankararlar/04-uzaktan-egitim-ve-yks-ertelenmesine-iliskin.pdf adresinden 05/09/2022 tarihinde alınmıştır.

[2] Yaman, İ. (2018). Türkiye'de İngilizce Öğrenmek: Zorluklar ve Fırsatlar . *RumeliDE Dil ve Edebiyat Araştırmaları Dergisi,* (11), 161-175. https://doi.org/10.29000/rumelide.417491

[3] İngilizce Yeterlik İndeksi [English Proficiency Index] (2020, Kasım). The world's largest ranking of countries and regions by English skills. https://www.ef.com.tr/epi/

[4] Sezer, A. (1988). Çağdaş Gelişmeler Işığında Türkiye'de Eğitim Fakültelerinin Yeri Ve Rolü. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 2, 183-189.

[5] Demircan, Ö. (1988). Dünden bugüne Türkiye'de yabancı dil. İstanbul: Remzi Kitabevi.

[6] Townsend, P., & Poh, H. J. (2008). An exploratory study of ınternational students studying and living in a regional area. Journal of Marketing for Higher Education, 18(2), 240-262. https://doi.org/10.1080/08841240802487411

[7] Mazzarol, T., & Soutar, G. N. (2001). The Global Market For Higher Education: Sustainable Competitive Strategies For The New Millennium. Edward Elgar Pblishing

[8] Karabulut D. (2021*). Hastane Bilgi Yönetim Sistemlerinde Veri Madenciliği: Hasta Profil Tahmini,* Yüksek Lisans Tezi, Karabük Üniversitesi,

[9] Can, Ş., Gerşil, M., (2019). Online Alışveriş Davranışlarının Satın Alma Niyetine Etkisinin Karar Ağacı ile Haritalandırılması*, Kesit Akademi Dergisi,* 5(21), 350-360. https://doi.org/10.29228/kesit.38836

[10] Yıldıztepe E., Kocataş A., (2018). Türkiye İşgücü Verilerinin Karar Ağacı Yöntemleriyle Analizi, *Çankırı Karatekin Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 8(2), 91-114. https://doi.org/10.18074/ckuiibfd.340236

[11] Aytekin Ç., Sütcü C., Özfidan U., (2018). Karar Ağacı Algoritması ile Metin Sınıflandırma: Müşteri Yorumları Örneği, *Uluslararası Sosyal Araştırmalar Dergisi*, 11(55), 782-792. http://dx.doi.org/10.17719/jisr.20185537249

[12] Emel G., Taşkın Ç., (2005). Veri Madenciliğinde Karar Ağaçları ve Bir Satış Analizi Uygulaması, *Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Dergisi* , 6(2), 221-236.

[13] Çakır E., Kamal B., (2021) . İstanbul Boğazı'ndaki Ticari Gemi Kazalarının Karar Ağacı Yöntemiyle Analizi, *Aquatic Research*, 4(1), 10-20. https://doi.org/10.3153/AR21002

[14] Kadirhanoğulları İ., Karadaş K., Özger Ö., Kadirhanoğulları M., (2021) . Karar Ağacı Algoritmatları ile Oragnik Ürün Tüketici Tercihlerinin Belirlenmesi : *Iğdır İli Örneği, Yüzüncü Yıl Üniversitesi Tarım Bilimleri Dergisi*, 31(1), 188-196. https://doi.org/10.3153/AR21002

[15] Sevüktekin, M. , Oğuzlar, A. , Aydın, B. & Nargeleçekenler, M. (2007). Karar Ağacı Yardımıyla Suçluların Özelliklerinin Belirlenmesi . *Öneri Dergisi* , 7 (27) , 291-298. https://doi.org/10.14783/maruoneri.686912

[16] Bardi Ş., Can A. V. (2021). Diskriminant Analizi Ve C5.0 Algoritması ile Finansal Başarısızlığın Tahmini: BİST Kobi Sanayi Endeksi'ndeki İşletmeler Örneği. *Ömer Halisdemir Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 14(3), 1071-1090. http://doi.org/.10.25287/ohuiibf.925344.

[17] Arslantürk Çöllü D., Akgün L., Eyduran E., (2020). Karar Ağacı Algoritmalarıyla Finansal Başarısızlık Tahmini: Dokuma, Giyim Eşyası ve Deri Sektörü Uygulaması*, Uluslararası Ekonomi ve Yenilik Dergisi*, 6(2), 225-246. https://doi.org/10.20979/ueyd.698738

[18] Koçak H.,(2020), Çalışan Örgütsel Bağlılıklarının Cart Karar Ağacı Algoritması ile Belirlenmesi, *Uluslararası İktisadi ve İdari Bilimler Dergisi,* 6(2), 67-87. https://doi.org/10.29131/uiibd.831032

[19] Yakut E., Korkmaz A., (2020), İnsani Gelişmişlik Endeksinin Karar Ağacı Algoritmaları ile Modellenmesi: BM' de Bir Uygulama 2010-2017 Dönemi, *Anadolu Üniversitesi Sosyal Bilimler Dergisi*, 20(2), 65-84. https://doi.org/10.18037/ausbd.758032

[20] Beşli N., Tenekeci M. E., (2020) . Uydu Verilerinden Karar Ağaçları Kullanarak Orman Yangını Tahmini, *Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi*, 11(3), 899-906, https://doi.org/10.24012/dumf.661925

[21] Okatan E., Işık A. H., (2020) , Sağlık Harcamalarının Tahmininde Karar Ağacının Kullanımı, *Mehmet Akif Ersoy Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 11(1), 86-94. https://doi.org/10.29048/makufebed.650463

[22] Demirel, Ş., Y. Giray, S. (2019). Karar Ağacı Algoritmaları ve Çocuk İşçiliği Üzerine Bir Uygulama, *Social Sciences Research Journal*, 8(4), 52-65

[23] Akbal E., Doğan Ş., Varol N., (2017) . Karar Ağaçları ile Telefon Dolandırıcılığı Verilerinin Analizi, Fırat Üniversitesi Fen ve Müh. Bil. Dergisi, 29(1), 171-177.

[24] Aksu Ç. M., Karaman E., (2017) , Karar Ağaçları ile Bir Web Sitesinde Link Analizi ve Tespiti, *Acta Infologıca*, 1(2), 84-91.

[25] Ersöz T., Özseven T., Ersöz F., (2017). Tüketicilerin Cep Telefonu Tercihlerinin Karar Ağacı ile Modellenmesi, *Gümüşhane Üniversitesi Sosyal Bilimler Enstitüsü Elektronik Dergisi*, 8(19), 129-136

[26] Pala, M. , Çimen, M. , Boyraz, Ö. , Yildiz, M. , Boz, A.. (2019). Meme Kanserinin Teşhis Edilmesinde Karar Ağacı Ve KNN Algoritmalarının Karşılaştırmalı Başarım Analizi. *Academic Perspective Procedia*, 2 (3), 544-552. https://doi.org/10.33793/acperpro.02.03.47

[27] Çalış A., Kayapınar S., Çetinyokuş T., (2014). Veri Madenciliğinde Karar Ağacı Algoritmaları İle Bilgisayar Ve İnternet Güvenliği Üzerine Bir Uygulama, *Endüstri Mühendisliği Dergisi*, 25(3), 2-19

[28] Çalış A., (2013) . *Veri Madenciliği Yaklaşımı İle Bireysel Müşterilerin Kredi Ödeme Performanslarının Değerlendirilmesi*, [Yüksek Lisans Tezi] Kocaeli Üniversitesi,

[29] Altunkaya H. İ., (2013). *Ülkelerin Uzun Dönem Kredi Notlarının Derecelendirilmesinde Önemli Değişkenlerin Veri Madenciliği Teknikleri Kullanılarak Belirlenmesi* [Yüksek Lisans Tezi] Hacettepe Üniversitesi

[30] Dolgun M. Ö., Ersel D., (2014). Doğrudan Pazarlama Stratejilerinin Belirlenmesinde Veri Madenciliği Yöntemlerinin Kullanımı, *İstatikçiler Dergisi:İstatistik&Aktüerya*, 7(1), 1-13

[31] Erbay, Ş. , Erdem, E. & Sağlamel, H. (2014). İyi Bir Yabancı Dil Öğretmen Profili: Özel Dil Kursları Yönetici Fikirlerinin Çapraz Mülakat Analizi . *Turkish Online Journal of Qualitative Inquiry* , 5 (4) ,41-61

[32] Özçelik, N. (2013). *Üniversite Öğrencilerinin İkinci Yabancı Dil Öğrenme Profili* , Erzincan Üniversitesi Eğitim Fakültesi Dergisi , 15 (1) , 123-147

[33] Köktürk T., (1997). *İlköğretim Okulları İkinci Kademe İngilizce Öğretmenlerinin Profili, Motivasyonu ve İş Tahmini* [Yüksek Lisans Tezi] Marmara Üniversitesi

[34] Özlüer Başer B.,Yangın M., Sarıdaş E.S., (2021). Makine Öğrenmesi Teknikleriyle Diyabet Hastalığının Sınıflandırılması, *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 25(1), 112-120. . https://doi.org/10.19113/sdufenbed.842460

[35] Özcan B., Turna C., (2021). Karar Ağaçları ile İnternet Alışverişlerinde Tüketiciyi Etkileyen Faktörlerin Analizi, *Journal Of Busıness In The Dıgıtal Age*, 4(2), 94-105. https://doi.org/10.46238/jobda.882832

[36] Büyükarıkan U. (2020). Finansal Performansa Etki Eden Finansal Değişkenlerin Chaid Karar Ağacıyla Belirlenmesi, Aydın İktisat Fakültesi Dergisi, 5(1), 1-10.

[37] Altun, M. (2017). *Veri Madenciliği ve Uygulama Alanları,* [Doktora Semineri Raporu] Akdeniz Üniversitesi

[38] Mecca, G., Raunıch, S., & Pappalardo, A. (2007). A new algorithm for clustering search results search results. *Data & Knowledge Engineering*, 62(3), 504-522. https://doi.org/10.1016/j.datak.2006.10.006

[39] Alpaydın, E. (2000). Zeki Veri Madenciliği: Ham Veriden Altın Bilgiye Ulaşma Yöntemleri. Bilişim 2000 Eğitim Semineri, Boğaziçi Üniversitesi

[40] Gargano, M.L. ve Raggad, B.G. (1999). Data mining–A Powerful Information Creating Tool, OCLC Systems & Services: International digital library perspectives, 15(2),81–90. https://doi.org/10.1108/10650759910276381

[41] Karabulut D. [2021]. *Hastane Bilgi Yönetim Sistemlerinde Veri Madenciliği: Hasta Profil Tahmini*, [Yüksek Lisans Tezi]Karabük Üniversitesi,

[42] Liao, S. (2003). Knowledge Management Technologies And Applications—Literature Review From 1995 To 2002, *Expert Systems With Applications* 2(25), 155–164. https://doi.org/ 10.1016/S0957-4174(03)00043-5

[43] https://www.oracle.com/tr/database/what-is-a-data-warehouse/ adresinden 14/02/2023 tarihinden alınmıştır.

[44] https://thinktech.stm.com.tr/uploads/docs/1608899913_stm-blog-derin-farklar-yapay-zeka-makine-ogrenmesi-ve-derin-ogrenme.pdf adresinden 10/10/2022 tarihinde alınmıştır.

[45] Han J., Kamber M.,Pei J., (2001). Data Mining Concepts and Techniques, Morgan Kaufman Publishers.

[46] Sezen. H. K., (2004). *Yöneylem Araştırması*, Ekin Kitapevi

[47] Nisbet R., Elder J., Miner G., (2009). *Handbook of Statistical Analysis and Data Mining Applications,* Academic Press

[48] Maimon, O., Rokach, L. (2010). *Data Mining And Knowledge Discovery Handbook*, Springer

[49] Baran Kılıçalan M., (2018*) Hanehalkı İşgücü Araştırma Verileri İle Veri Madenciliği Yöntemlerinin Uygulanması ve Modellerin Karşılaştırılması* [Yüksek Lisans Tezi] Hacettepe Üniversitesi

[50] Bahety, A. (2014). Extension and Evaluation of ID3–Decision Tree Algorithm.

[51] Adhatrao, K., Gaykar, A., Dhawan, A., Jha, R., & Honrao, V. (2013). Predicting Students' Performance using ID3 and C4. 5 Classification Algorithms, *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, 3(5),39-52, DOI:10.5121/ijdkp.2013.350465

[52] Atılgan, E. (2011). *Karayollarında Meydana Gelen Trafik Kazalarının Karar Ağaçları ve Birliktelik Analizi İle İncelenmesi* [Yayınlanmamış Yüksek Lisans Tezi] Hacettepe Üniversitesi

[53] Akpınar, H. (2000). Veri Tabanlarında Bilgi Kesfi ve Veri Madenciliği, *İ.Ü. İşletme Fakültesi Dergisi*, 29(1), 1-22.

[54] Arslan, H., (2008) *Sakarya Üniversitesi Web Sitesi Erişim Kayıtlarının Web Madenciliği İle Analizi* [Yüksek Lisans Tezi] Sakarya Üniversitesi

[55] Seker S.E., Çankır B., Arslan M.L, (2014) Information and Communication Technology Reputation for XU030 Quote Companies, *International Journal of Innovation and Technology Management*, 5(3),221-225

[56] Center for Media Justice, Color of Change ve Sum of Us, 2013: 19

[57] Can, Ş. (2017*). Veri Madenciliği ve Eğitim Sektöründe Bir Uygulama.* [Yayınlanmamış Yüksek Lisans Tezi] Manisa Celal Bayar Üniversitesi

ÖZGEÇMİŞ

Ad-Soyad :Sevim Şevval ZOROĞLU

ÖĞRENİM DURUMU:

- **Lisans :** 2019, Sakarya Üniversitesi, Mühendislik Fakültesi , Endüstri Mühendisliği Bölümü
- **Yükseklisans :** 2023, Sakarya Üniversitesi , Endüstri Mühendisliği Anabilim Dalı, Endüstri Mühendisliği Programı

MESLEKİ DENEYİM VE ÖDÜLLER:

- Eylül 2018- Ocak 2019 yılları arasında Yazaki Otomotiv Firmasında Aday Mühendis olarak çalıştı.
- 2022 Nisan ayından beri AKÇELİK GROUP firmasında Üretim Planlama Mühendisi olarak çalışmaktadır.

TEZDEN TÜRETİLEN ESERLER:

DİĞER ESERLER:

Zoroglu, S.S., Yalciner Yılmaz A., (2022). Determining the Priority Criteria for Personnel Selection With Fuzzy Dematel and Grey-Based Dematel Approaches, Recent Advances in Intelligent Manufacturing and Service Systems, Springer Singapore