KOCAELİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLİŞİM SİSTEMLERİ MÜHENDİSLİĞİ ANABİLİM DALI

YÜKSEK LİSANS TEZİ

TIP FAKÜLTESİ ÖĞRENCİLERİNİN KURUL SINAVI BAŞARILARININ VERİ MADENCİLİĞİ ALGORİTMALARI KULLANILARAK İNCELENMESİ

ERGÜL MADEN

KOCAELİ 2021

KOCAELİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

BİLİŞİM SİSTEMLERİ MÜHENDİSLİĞİ

ANABİLİM DALI

YÜKSEK LİSANS TEZİ

TIP FAKÜLTESİ ÖĞRENCİLERİNİN KURUL SINAVI BAŞARILARININ VERİ MADENCİLİĞİ ALGORİTMALARI KULLANILARAK İNCELENMESİ

ERGÜL MADEN

Prof. Dr. Mehmet YILDIRIM		
Danışman,	**Kocaeli Üniv.**	**.....**
Doç. Dr. Serdar SOLAK		
Jüri Üyesi,	**Kocaeli Üniv.**	**.....**
Dr. Öğr. Üyesi Adem TUNCER		
Jüri Üyesi,	**Yalova Üniv.**	**.....**

Tezin Savunulduğu Tarih: 17.06.2021

ÖNSÖZ VE TEŞEKKÜR

Bu çalışmada Tıp Fakültesi öğrencilerinin kurul sınav sonuçlarının veri madenciliği yöntemleri kullanılarak incelenmesi amaçlanmıştır. Sonuç olarak eğitim öğretim süreçlerinde oluşan veri kümeleri kullanılarak elde edilen anlamlı bilgi, eğitim öğretim süreçlerinin kalite ve verimliliğinin artırılması amacı ile kullanılabilcektir.

Yüksek lisans eğitimimim süresince bilgisi ve tecrübesi ile her zaman yanımda olan, bana yol gösteren ve yoğun çalışma temposuna rağmen desteğini hiçbir zaman esirgemeyen saygıdeğer hocam Prof. Dr. Mehmet YILDIRIM'a sonsuz sevgi ve teşekkürlerimi iletmek isterim.

Çalışmam boyunca yardımlarını ve desteklerini esirgemeyen değerli dostlarım Öğr. Gör. Samet DİRİ, Müh. Murat UZUN, Müh. Koray DUMAN ve çalışma arkadaşlarıma, maddi ve manevi desteklerini hiçbir zaman esirgemeyen sevgili annem Selver MADEN, babam Erdoğan MADEN'e sonsuz sevgi ve minnettarlıkla teşekkür ederim.

Ayrıca tanıştığımız günden beri desteği ile beni güçlendiren hayat arkadaşım, sevgili eşim Funda ALAÇAM MADEN'e ve oğullarım Erdem Ali MADEN ve Görkem Ege MADEN'e teşekkür ederim.

Mayıs – 2021 Ergül MADEN

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

SİMGELER VE KISALTMALAR DİZİNİ

Kısaltmalar

CRISP-DM : Cross Industry Standart Process for Data Mining
FN : False Negative
FP : False Positive
KDD : Knowledge Data Discovery
KNIME : Konstanz Information Miner
MSSQL : Microsoft Structured Query Language (Microsoft Yapılandırılmış Sorgulama Dili)
SEMMA : Sample Explore Modify Model Evalute
TN : True Negative
TP : True Positive

TIP FAKÜLTESİ ÖĞRENCİLERİNİN KURUL SINAVI BAŞARILARININ VERİ MADENCİLİĞİ ALGORİTMALARI KULLANILARAK İNCELENMESİ

ÖZET

Birçok sektörde bilişim teknolojilerinde yaşanan olumlu gelişmelere paralel olarak dijital ortamda oluşan ve depolanan veri miktarı her geçen gün artmaktadır. Bu durum, oluşan verilerin analiz edilerek veri içerisindeki saklı bilginin ortaya çıkarılması ve değere dönüştürülmesi gereksinimini beraberinde getirmiştir. Tüm bu gelişmeler neticesinde, büyük veri yığınları içerisindeki saklı bilgi ve örüntülerin keşfedilmesi olarak adlandırılan veri madenciliği süreci önemini ve kullanım alanlarını her geçen gün artırmaktadır.

Bu tez çalışmasında, tıp fakültesi kurul sınavları sonucu oluşan veri kümeleri üzerinde veri madenciliği sınıflandırma yöntemleri uygulanarak, veri kümesi içinde yer alan saklı bilgi ve örüntülere ulaşılması hedeflenmiştir. Süreç sonucunda elde edilen saklı bilgi ve örüntülerden, eğitim öğretim süreçlerinin planlanması ve karar süreçlerinde faydalanılarak, eğitim süreçlerinin kalite ve verimliliğine olumlu yönde katkılar sağlaması amaçlanmıştır.

Çalışma kapsamında; Karar Ağaçları, Yapay Sinir Ağları, Rastgele Orman, Naive Bayes ve K-En Yakın komşuluk veri madenciliği yöntemleri kullanılmış ve başarı oranları karşılaştırılmıştır. Tıp fakültesi kurul sınavları sonucu oluşan veri kümeleri üzerinde veri madenciliği sınıflandırma yöntemleri uygulanarak, öğrencilerin akademik başarı durumlarının erken dönemlerde tahmin edilebileceği bir model geliştirilmiştir.

Anahtar kelimeler: KNIME, Phyton, Tıp Eğitimi, Veri Madenciliği, Veri Ön İşleme.

ANALYSIS OF FACULTY OF MEDICINE STUDENTS' SUCCESS OF THE BOARD EXAM BY USING DATA MINING ALGORITHMS

ABSTRACT

In parallel with the positive developments in information technologies, the amount of data generated and stored in digital media in many sectors is increasing every day. This situation has brought the necessity of analyzing the data, revealing the hidden information in the data and transforming it into value. As a result of these developments, the importance and usage areas of the data mining process, which is defined as the discovery of hidden information and patterns in large data stacks, are increasing day by day.

In this thesis, it was aimed to reveal hidden information and patterns by applying data mining classification methods on the data sets formed as a result of the medical faculty board exams. It is aimed to improve the quality and efficiency by making use of these hidden information and patterns in the planning of educational processes and decision-making processes.

Within the scope of the study, Decision Trees, Artificial Neural Networks, Random Forest, Naive Bayes and K-Nearest neighborhood data mining methods were used and their success rates were compared. By applying data mining classification methods on the data sets formed as a result of the medical faculty board exams, a model has been developed in which the academic success of the students can be predicted in the early stages.

Keywords: KNIME, Phyton, Medicine Education, Data Mining, Data Preprocessing.

GİRİŞ

Her geçen gün daha fazla dijitalleşen dünyada atılan her adım, geride bir veri izi bırakmakta ve bu durum veri miktarının çok ciddi oranlarda artışına neden olmaktadır. Bilişim teknolojilerinde yaşanan donanımsal ve yazılımsal gelişmeler ve genişbant internet kullanımının yaygınlaşması, her geçen gün daha fazla sayıda kullanıcının internete erişim imkanı bulmasını sağlamıştır. 2020 yılında, 50 milyardan daha fazla cihazın internete bağlı olduğu tahmin edilmektedir. Her geçen gün çok daha fazla sayıda kullanıcının internete erişim imkânı bulması ve veri depolama teknolojilerinde yaşanan olumlu gelişmeler neticesinde, geçmişe kıyasla çok büyük boyutlarda verinin üretilmesi ve depolanması, çok daha kolay hale gelmiştir (Marr, 2017, Yılmaz, 2015). Büyük boyutlardaki veri kümelerinin oluştuğu ve depolandığı alanlardan bir tanesi de eğitim-öğretim faaliyetleridir.

Eğitim ve öğretim faaliyetlerinin kalite ve verimliliği, günümüz toplumlarının gelişmişlik düzeylerini belirleyen en önemli etkenler arasındadır. Eğitimde kalite ve verimliliğin en üst düzeye çıkartılması, toplumların rekabet gücünü artırmaktadır. Bu kapsamda, eğitim faaliyetleri esnasında oluşan verilerin değerlendirilerek, ortaya çıkan anlamlı bilgilerin karar süreçlerinde kullanılması, eğitim faliyetlerinin kalitesine olumlu katkılar sağlayacaktır.

Farklı sektörlerde oluşan verilerin işlenerek anlamlı bilgiye ulaşılmasında geleneksel yöntemler yetersiz kalmaktadır (Garcia ve diğ., 2016). Bu durum veri madenciliği kavramının ortaya çıkışına neden olmuştur (Demir vd., 2020). Veri madenciliği, farklı yöntemler ile farklı kaynaklardan elde edilen veriler üzerinde işlemler yapılarak, büyük veri kümelerinde gizlenmiş anlamlı bilgi ve örüntüleri keşfetme sürecidir (Şeker, 2013). Bu süreçte, bilimsel problem çözme teknikleri, matematik, istatistik ve yazılım geliştirme disiplinleri başta olmak üzere, birçok farklı disiplinden faydalanılmaktadır (Oğuzlar, 2003).

Bu tez çalışmasında, tıp fakültesi öğrencilerinin kurul derslerine ait sınav sonuç verileri kullanılarak, öğrencilerin akademik başarı ve başarısızlıklarının erken

dönemlerde tahmin edilmesi amaçlanmıştır. Kullanılan veri madenciliği yöntemlerinin başarı oranları değerlendirilmiştir. Gerçekleştirilen çalışma sonucunda elde edilen anlamlı bilginin, karar süreçlerinde kullanılarak, eğitim faaliyetlerinin kalite ve verimliliğine olumlu katkılar sağlanması amaçlanmıştır. Veri madenciliği uygulamalarında sürecin %80'lik kısmını veri hazırlama ve veri ön işleme aşaması oluşturmaktadır. Kaliteli bilgiye sahip olmanın ön şartı, kaliteli veriye sahip olmaktır. Bu nedenle veri hazırlığı/ön işleme, veri madenciliği sürecinin en önemli adımıdır (Piramuthu, 2003). Veri madenciliği süreçlerinin en önemli aşamalarından olan veri ön işleme aşamasında, eksik veri probleminin çözümüne dönük çalışmalar gerçekleştirilmiştir. Eksik veri probleminin çözümünde farklı yöntemler kullanılarak, kullanılan yöntemlerin oluşturulan veri madenciliği modellerinin başarısına etkisi araştırılmıştır.

Veri madenciliği yöntemleri kullanılarak, farklı veri kaynaklarından elde edilmiş olan veriler üzerinde, akademik başarı ve başarısızlığın modellenmesi üzerine gerçekleştirilmiş olan, yedi adet çalışma incelenmiş ve aşağıda yer alan değerlendirmelere ulaşılmıştır.

Bırtıl (2011) yüksek lisans tezinde, lise öğrencilerine uygulanan anket verilerini kullanarak, öğrencilerin akademik başarısızlık nedenlerinin tespit edilmesini amaçlamıştır. Çalışmada kümeleme algoritmaları kullanılmış ve öğrencilerin üç farklı kümeye ayrıştığı gözlenmiştir. Çalışma sonucunda tespit edilen ve akademik başarısızlığa yol açan nedenlere yönelik iyileştirmeler yapılarak, oluşturulan veri madenciliği modelinin tekrar uygulanması önerilmiştir.

Şengür (2013) yüksek lisans tezinde, veri madenciliği yöntemlerinden olan Yapay Sinir Ağları ve Karar Ağaçlarını kullanarak Fırat Üniversitesi, Eğitim Fakültesi, Bilgisayar ve Öğretim Teknolojileri Eğitimi Bölümü (BÖTE) öğrencilerinin mezuniyet notlarını tahmin etmeye çalışmıştır. Çalışmada 127 adet öğrencinin eğitimleri süresince aldıkları derslere ait sınav başarı notları kullanılmıştır. İki farklı senaryo oluşturulmuştur. Birinci senaryoda ilk iki yıla ait başarı notları kullanılarak mezuniyet notlarının tahmini amaçlanmış, ikinci senaryoda ise ilk üç yıla ait başarı notları kullanılarak öğrencilerin mezuniyet notlarının tahmin edilmesi hedeflenmiştir.

Yapılan çalışmada Yapay Sinir Ağları yönteminin, karar ağaçları yöntemine oranla daha yüksek başarım oranına sahip olduğu tespit edilmiştir.

Akçapınar (2014) doktora tezinde, çevrimiçi öğrenme ortamındaki öğrenci verilerini kullanmıştır. Farklı veri madenciliği algoritmaları kullanılarak bir model geliştirilmiş, öğrencilerin erken dönemlerde akademik performanslarının tahmini ve verilerin analiz edilmesini amaçlamıştır. Çalışmada kullanılan veri madenciliği algoritmalarının başarı oranları karşılaştırılmıştır. Ayrıca veri ön işleme ve öznitelik seçme tekniklerinin, kullanılan veri madenciliği algoritalarının başarısına etkisi gözlemlenmiştir. Çalışma sonucunda, oluşturulan model ile öğrenci başarılarının erken dönemlerde tahmin edilebileceği ve öğrencilerin çevrim içi öğrenme ortamında geçirdikleri süre ve etkileşimin, öğrenci başarısına doğrudan etki ettiği sonucuna ulaşılmıştır.

Özdemir (2016) doktora tezinde, farklı sosyo-demoğrafik özelliklere sahip lise öğrencilerinin verileri üzerinde, CRISP-EDM (Cross Industry Standard Process for Educational Data Mining) süreç modelini kullanarak, farklı veri madenciliği algoritmaları ile modeller oluşturmuştur. Yapılan analizlerde akademik başarının tahmin edilmesinde, C4.5 karar ağacı algoritmasının daha doğru sonuçlar verdiği tespit edilmiştir.

Buluz (2017) yüksek lisans tezinde, öğrencilerin demografik özellikleri ve geçmiş dönemlere ait akademik başarıları üzerinde, sınıflandırma metotlarından olan Naive Bayes ve K-En Yakın Komşuluk yöntemi kullanarak oluşturulan modellerin performanslarını incelemiştir. Daha sonra, öğrencilere ait demoğrafik ve akademik veriler çizgelerle ifade edilerek, öğrencilere ait özellikleri ifade eden yeni örüntüler keşfedilmiştir. Keşfedilen yeni örüntüler ile veri kümesi zenginleştirilmiştir. Oluşan yeni veri kümesi, veri madenciliği modeline tekrar uygulandığında, kullanılan yöntemlerin başarı oranlarının artış gösterdiği gözlenmiştir.

Aydemir (2017) yüksek lisans tezinde, meslek yüksek okulu öğrencilerine ait verileri kullanarak, öğrencilerin akademik başarılarının tahminini amaçlamıştır. Çalışmada sınıflandırma algoritmaları kullanılarak, oluşturulan veri madenciliği modellerinde kullanılan algoritmalar içerisinde, en yüksek başarı oranına sahip algoritma belirlenmeye çalışılmıştır. 1387 öğrencinin verileri kullanılarak gerçekleştirilen çalışmada, not ortalaması bağımlı değişkeni kullanılarak yapılan başarım tahmininde

en iyi sonucu Sıralı Minimum Optimizasyon (SMO) algoritmasının, mezuniyet yılına göre başarım tahmini yapıldığında en iyi sonucu J4.8 ve Naive Bayes algoritmalarının verdiği gözlemlenmiştir.

Altun (2019) doktora tezinde, belirli tarih aralıklarında farklı bölümlerden mezun olan 3773 öğrencinin sınav sonuçları, dönem sonu notları, dönemlere ait ortalamaları, mezuniyet ortalamaları, farklı kaynaklardan derlenerek toplanmıştır. Toplanan veriler üzerinde yapay sinir ağları ve çoklu doğrusal regresyon algoritmaları ile modeller geliştirilmiş ve kullanılan algoritmaların başarı performansları değerlendirilmiştir. Geliştirilen iki modelde öğrencilerin erken dönemlere ait başarı notları kullanılarak, öğrencilerin ilerleyen dönemlerdeki başarı durumlarını kestiren bir model geliştirilmiştir.

Bu tez çalışması 4 bölümden oluşmaktadır.

Birinci bölümde; veri madenciliğinin genel bir tanımı yapılmış, kullanım alanları ve tarihsel gelişimi hakkında bilgiler verilerek, uygulamada kullanılacak olan CRISP-DM veri madenciliği süreci ayrıntılı olarak açıklanmıştır. Model doğrulama yöntemleri ve model performans analiz yöntemlerinden bahsedilmiştir.

İkinci bölümde; çalışmamızda kullanılan veri madenciliği sınıflandırma yöntemleri ile ilgili ayrıntılı bilgiler sunulmaktadır. Ayrıca, tahmin edici veri madenciliği yöntemlerinden olan kümeleme ve birliktelik kuralı yöntemlerine değinilmiştir.

Üçüncü bölümde; gerçekleştirilen uygulama süreci ayrıntılı olarak ele alınmıştır. Uygulama geliştirme ortamı tanıtılmıştır. Veri setinin toplanması, depolanması, veri analiz ve veri ön işleme süreci, veri madenciliği modelinin oluşturulması ve değerlendirilmesi aşamaları ayrıntılı olarak açıklanmıştır. Veri ön işleme sürecinde kullanılan yöntemler açıklanmış ve oluşturulan modelin performans analizleri gerçekleştirilmiştir.

Dördüncü bölümde; çalışmanın eğitim-öğretim süreçlerine sağlayacağı katkılar ve bu alanda yapılabilecek çalışmalar ile ilgili önerilerden bahsedilmiştir.

1. VERİ MADENCİLİĞİ

Her geçen gün çok daha büyük boyutlarda verinin üretildiği bir dünyada yaşıyoruz. Bu durum oluşan verilerin en hızlı ve en doğru şekilde analiz edilerek, değere dönüştürülmesi gerekliliğini ortaya çıkarmıştır. Bu durumun kaçınılmaz bir sonucu olarak, veri madenciliği kavramı ortaya çıkmıştır (Han vd., 2011). Veri madenciliği, veri kümeleri üzerinde matematiksel ve istatistiksel teknikler kullanılarak, veri kümesi içinde yer alan potansiyel saklı bilgi ve örüntülere ulaşma sürecidir (Akküçük, 2011). En genel ifade ile veriden bilgiye ulaşmayı amaçlayan süreçtir (Şeker, 2013). Veri madenciliği eğitim, bankacılık ve finans, sağlık, mühendislik, savunma sanayi ve satış-pazarlama başta olmak üzere, birçok farklı alana ait veri kümeleri kullanılarak, eldeki ham veriden saklı bilgiye ulaşmayı amaçlayan multidisipliner bir süreçtir. Politikacılar için seçmen eğilimlerinin belirlenmesi, eğitimciler için akademik başarının modellenmesi, finans ve bankacılık sektörü için kredi risk analizinin gerçekleştirilmesi gibi birçok alanda çözümler sunmaktadır (Aggarwal, 2015).

Şekil 1.1. Veri madenciliği ile ilişkili disiplinler (Savaş vd., 2012)

Veri madenciliği süreçlerinde birçok farklı disiplinden, doğrudan veya dolaylı olarak faydalanılmaktadır. Veri madenciliği süreci İstatistik, Makine Öğrenmesi, Yapay Zekâ, Veritabanı Yönetimi gibi farklı disiplinlerin yoğun şekilde kullanıldığı bir süreçtir (Altay, 2019).

1.1. Veri Madenciliği Tarihsel Gelişimi

Veri madenciliği tarihsel gelişim sürecinde başlangıç noktasını, ilk sayısal bilgisayar olan ENIAC (Electrical Numerical Integrator And Calculator)'ın keşfi olarak kabul edebiliriz. İlk bilgisayardan günümüze kullanıcı ve sektörel ihtiyaçlar doğrultusunda, donanımsal ve yazılımsal olarak çok büyük gelişimler kaydedilmiştir (Savaş vd., 2012). Bu gelişmeler neticesinde veri madenciliği kavramı önemli değişimlere uğramıştır. Veri madenciliği tarihsel gelişimi Şekil 1.2'de özetlenmiştir. Veri madenciliği tarihsel gelişimi 4 dönemde incelenebilir (Han, Pei ve Kamber, 2011)

Gelişim Adımları	Cevaplanan Karar Problemi	Kullanılabilen Teknolojiler	Ürün Sağlayıcıları	Karakteristikler
Veri Toplama (1960'lar)	"Benim toplam karım geçen 5 yılda ne kadardı?"	Bilgisayarlar, Teypler, Diskler	IBM, CDC	Geriye dönük, statik veri dağıtımı
Veri Erişimi (1980'ler)	"İngiltere'de geçen mart ayında birim satışları ne kadardı?"	İlişkisel Veri tabanları, SQL, OBDC	Oracle, Sybase, Informix, IBM, Microsoft	Kayıt düzeyinde geriye dönük, dinamik veri dağıtımı
Veri Ambarlama ve Karar Destek Sistemleri (1990'lar)	"İngiltere'de geçen mart ayında birim satışları ne kadardı? Boston'a aşağı inebilecek"	OLAP, Çok Boyutlu Veri tabanı Sistemleri, Veri ambarları	Pilot, Comshare, Arbor, Cognos, Microstrategy	Çoklu düzeylerde, geriye dönük dinamik veri dağıtımı
Veri Madenciliği (Bugün)	"Gelecek ay Boston'daki birim satışlar muhtemelen ne olabilir? Neden?"	İleri düzeyde algoritmalar, çok işlemcili bilgisayarlar, büyük veri tabanları	Pilot, Lockheed, IBM, SGI, SPSS, SAS, Microsoft, vs.	Geleceğe dönük, proaktif enformasyon dağıtımı

Şekil 1.2. Veri madenciliği tarihsel gelişimi (Aldana, 2000)

1960'lar, veri tabanı ve veri depolama kavramlarının ortaya çıktığı ve yaygınlaşmaya başladığı dönemdir.

1980'ler, ilişkisel veri tabanı yönetim sistemlerinin ortaya çıktığı ve yaygınlaşmaya başladığı dönemdir. Bu dönemde, veri analiz teknikleri kullanılmaya başlanmıştır.

1990'lar, internet kullanımının artması ve web tabanlı veri tabanlarının yaygınlaşması ile veri tabanlarında depolanan veri miktarının ciddi oranda artış gösterdiği dönemdir. Veri madenciliği temel kavramlarının ortaya çıkmaya başladığı ve ilk veri madenciliği yazılımının gerçekleştirildiği dönemdir.

2000'ler, veri depolama ortamları, veri toplama ekipmanları ve işlemci hızlarının gelişmesine parelel olarak, veri madenciliği çalışmalarında ciddi gelişimlerin yaşandığı ve birçok farklı alanda veri madenciliği uygulamalarının geliştirildiği dönemdir. Sosyal ağlar, bulut bilişim, nesnelerin interneti gibi kavramların ortaya çıkışı, üretilen veri miktarında çok ciddi artışlara neden olmuş ve bu durum veri madenciliğine olan ilgi ve yönelimi artırmıştır (Han, Pei ve Kamber, 2011, Savaş vd., 2012).

Veri madenciliği tarihsel gelişimi incelendiğinde, donanımsal ve yazılımsal gelişmelere paralel olarak, özellikle veri depolama teknolojilerinin hız ve kapasite olarak gelişmesi neticesinde, veri madenciliği süreçlerinin etki ve öneminin arttığı, uygulama alanlarının genişlediği gözlenmektedir.

1.2. Veri Madenciliği Uygulama Alanları

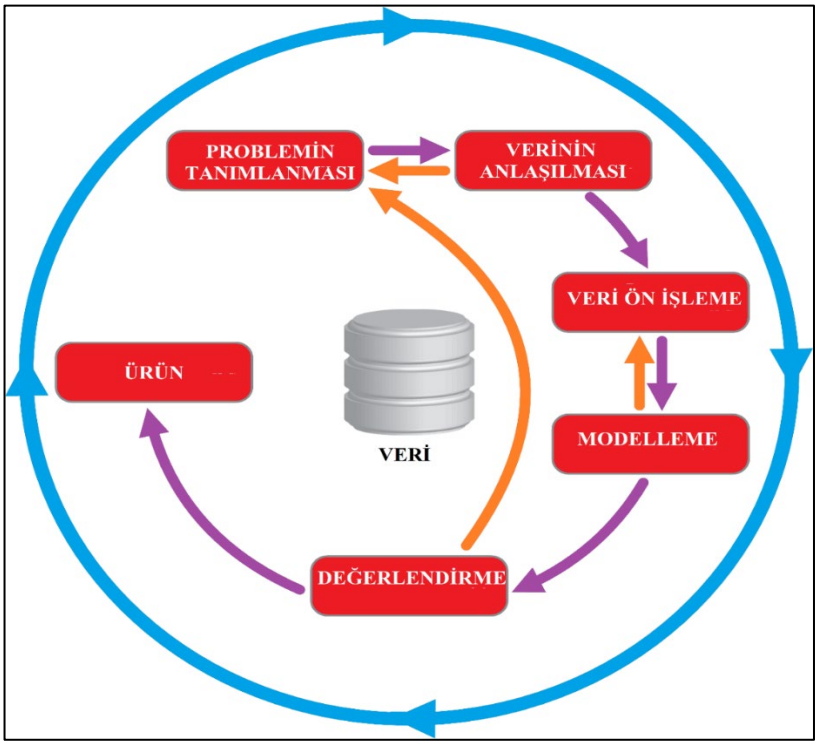
Veri madenciliği uygulama alanlarını, en genel ifade ile verinin olduğu her yer olarak tanımlayabiliriz. Bugün, perakende sektöründen sosyal ağlara, teknoloji şirketlerinden hükümetlere, medya sektöründen finans kuruluşlarına kadar, pek çok farklı alanda faaliyetlerini sürdüren kurum ve kuruluş, faaliyetleri esnasında ortaya çıkan veri kümeleri üzerinde, veri madenciliği süreçlerinden faydalanarak anlamlı bilgiye ulaşmaktadır (Marr, 2017). Pazarlama alanında müşteri ve satış analizi, mevcut müşterilerin korunarak yeni müşterilerin kazanılması ve satış tahmini yapılması için kullanılır. Bankacılık ve finans sektörü için, kredi risk analizi, dolandırıcılık tespiti, sigortacılık sektöründe ise riskli müşterilerin tespiti için kullanılır. Sağlık sektöründe, hastalık teşhisi ve hastaya özgü tedavi yöntemlerinin belirlenmesi, çözümünde veri madenciliği yöntemlerinin kullanıldığı problemlerden bazılarıdır (Savaş vd., 2012).

Örneğin, 2 Milyon 100 bin çalışanı, 11 binden fazla mağazası ile dünyanın 15 farklı ülkesinde faaliyetlerini sürdüren, dünyanın en büyük perakende şirketi WALLMART, elde ettiği ham verileri etkin biçimde değere dönüştürmektedir. Müşteri etkileşimi

esnasında oluşan gerçek zamanlı verilerin yanı sıra, meteorolojik veriler, ekonomik veriler ve sosyal medya verilerini kullanarak, faaliyetleri esnasında oluşabilecek hata ve aksaklıkların saptanması ve çözümü için geçen süreyi, iki ila üç haftadan 20 dakikaya kadar düşürmüştür (Marr, 2016). Bu durum, kurumların hızlı ve sağlıklı kararlar almasında, veri madenciliği süreçleri sonucunda ortaya çıkan anlamlı bilginin önemini ortaya koymaktadır.

1.3. Veri Madenciliği Süreci

Veri madenciliği süreçlerinde, yaygın olarak üç temel yöntem kullanılmaktadır. Bunlar, SAS firması tarafından geliştirilen SEMMA (Sample Explore Modify Model Evalute), KDD (Knowledge Data Discovery), ve CRISP-DM (Cross Industry Standart Process for Data Mining)'dir (Şeker, 2018). Bu tez çalışmasında, CRISP-DM veri madenciliği süreç modeli kullanılmıştır. CRISP-DM, kullanıldığı sektörden ve kullanılan yazılımdan tamamen bağımsız bir veri madenciliği süreç modelidir. CRISP-DM, veri madenciliği süreçlerini standartlaştırmaktadır (Garcia, 2016). CRISP-DM veri madenciliği metodolojisi hiyerarşik bir süreçtir ve altı aşamadan oluşmaktadır (Şeker, 2018, Chapman vd., 2000). CRISP-DM veri madenciliği metodolojisi Şekil-1.3'de gösterilmiş ve süreç adımları aşağıda açıklanmıştır.



Şekil 1.3. CRİSP-DM veri madenciliği süreci (Şeker, 2018)

1.3.1. Problemin tanımlanması

CRISP-DM veri madenciliği süreç modeli, problemin tanımlanması adımı ile başlar. Sürecin sağlıklı işleyebilmesi için, bu adımda çözülmesi planlanan problemin, bir veri madenciliği problemine indirgenmesi gerekmektedir. Bu aşamada, süreç sonunda ulaşılması planlanan hedef ve kazanımlar net bir biçimde ortaya konulmalı, problemin çözümüne yönelik gereksinimler belirlenmelidir (Chapman vd., 2000). Bu süreç adımınında yaşanacak planlama ve değerlendirme hataları, sürecin genelinin başarısına doğrudan etki etmektedir (Kıyak, 2006). Problemin tanımlanması adımı, iş amaçlarının belirlenmesi, mevcut durumun değerlendirilmesi, veri madenciliği amaçlarının belirlenmesi ve proje planının oluşturulması alt adımlarından oluşmaktadır (Chapman vd., 2000). Problemin tanımlanması adımından sonra, CRISP-DM süreç modeli, verinin anlaşılması adımı ile devam etmektedir.

1.3.2. Verinin anlaşılması

Verinin anlaşılması adımı, verinin toplanması, tasnif edilmesi ve incelenmesi işlemlerinden oluşur. Bu aşamada, veriye herhangi bir müdahalede bulunulmaz. Uygulamada kullanılacak veriler, farklı kaynaklarda yer alıyorsa, bu aşamada bir araya getirilerek veri seti oluşturulur. Veri ön işleme aşamasında yapılacak çalışmalara dönük olarak, bu aşamada veri seti analiz edilir ve veri setindeki problemler tespit edilir. Veri seti içerisinde kirli, gürültülü veya eksik veri problemlerine dönük tespitler gerçekleştirilir (Şeker, 2018). Verinin doğru analiz edilmesi, problemin çözümü için en doğru veri setinin belirlenerek, zaman ve maliyet artışının önüne geçilmesi açısından önemlidir. Verinin anlaşılması sürecinde, veri görselleştirme tekniklerinden faydalanılabilir. Verinin anlaşılması aşamasından sonra süreç, veri ön işleme aşaması ile devam etmektedir.

1.3.3. Veri ön işleme

Bu aşama, veri madenciliği sürecinin başarısında çok büyük belirleyiciliğe sahiptir. Ancak hak ettiği ölçüde nadiren araştırılır (Piramuthu, 2003). Bu aşamada veri seti, model oluşturma aşaması için düzenlenir. Veri ön işleme sürecinin başarısı, oluşturulan modelin ve veri madenciliği sürecinin başarısına ve çalışma zamanına doğrudan etki etmektedir (Aggarwal, 2015). Başarılı bir veri madenciliği sürecinin en

temel gereksinimi, kaliteli veridir. Veri ön işleme aşaması; veri seçimi, veri temizleme, veri standardizasyonu, veri indirgeme ve değişken dönüşümü gibi aşamalarından oluşmaktadır (Chapman vd., 2000). Kullanılacak olan veri setinde tespit edilen problemlere göre, belirtilen işlemlerin tamamı veya bir kısmı uygulanabilir. Veri üzerinde gerçekleştirilecek veri ön işleme adımlarının belirlenmesinde, verinin anlaşılması sürecinde gerçekleştirilen veri analizi neticesinde elde edilen bulgular, izlenecek stratejinin tespit edilmesinde belirleyicidir (Garcia, 2016). Veri setinde karşılaşılabilecek sorunlar ve çözüm önerileri aşağıda açıklanmıştır.

1.3.3.1. Veri temizleme

Gerçek dünya verileri gürültülü, eksik veya aykırı değerler içerebilir. Bu durum, veri madenciliği sürecinin başarısını olumsuz yönde etkiler (Han vd., 2011). Veri temizleme aşamasında, veri setinde yer alan gürültülü, eksik ve aykırı gözlem problemlerinin çözümüne dönük çalışmalar gerçekleştirilir.

Eksik veri, veri setinde makine veya insan hataları sonucu bazı verilerin eksik olması durumunu ifade etmektedir. Veri setinde yer alan gözlem değerlerinin boş veya NULL olması eksik veriye örnektir. Eksik veri problemine dönük çözüm yöntemleri aşağıda açıklanmıştır (Bozkır, 2009).

- Eksik veri içeren kayıtlar ihmal edilerek silinebilir. Silinen kayıt sayısının fazla olması oluşturulan veri madenciliği modelinin başarısını olumsuz yönde etkileyebilir.
- Eksik veri içeren kayıtlar elle doldurulabilir. Zaman ve maliyet açısından dezavantajlı bir yöntemdir. Büyük boyutlu veri kümeleri için mümkün olmayabilir.
- Eksik veri içeren kayıtlara sabit bir değer atanabilir. Örneğin, öğrenim öğrenim durumu alanı boş olan kayıtlar için “Bilinmiyor” değeri girilebilir.
- Ortalama değer yazılabilir. Örneğin yaş alanı boş kayıtlar için veri setindeki diğer kayıtlara ait yaş değerlerinin aritmetik ortalaması alınarak, yaş değeri boş olan alanlara elde edilen değer girilebilir.
- Eksik veri içeren kayıtlar, mevcut veri seti üzerinde makine öğrenmesi algoritmaları kullanılarak tamamlanabilir.

Aykırı gözlem ve gürültülü veri, makine veya insan hataları sonucu veride ortaya çıkan bozulmalardır. Örneğin veri setinde insana ait yaş değişkeni alanında -500 değerinin yer alması gürültülü veriye örnek olarak verilebilir. Gürültülü ve aykırı veri problemine dönük çözüm yöntemleri aşağıda açıklanmıştır (Bozkır, 2009).

- Kutulama yöntemi

Gürültülü veri problemleminin çözümü için kullanılmaktadır. Veriler küçükten büyüğe veya büyükten küçüğe her kutuda eşit sıklıkta veri olacak şekilde kutulara ayrılır. Oluşan kutularda yer alan değerler, o kutuda yer alan değerlerin ortalama, medyan veya alt-üst sınır değerlerine kullanılarak düzenlenir. Kutulama yöntemine ait adımlar 9,15,22,5,35,29,22,26,25 örnek veri setinden faydalanılarak aşağıda açıklanmıştır.

Adım1: Veriler küçükten büyüğe sıralanır.

Veri Seti: 9,15,22,5,35,29,22,26,25

Sıranlamış Veri Seti: 5,9,15,22,22,25,26,29,35

Adım2: Veriler eşit derinlikli kutulara ayrılır.

Adım3: Veriler ortalama, medyan veya alt-üst sınır değerlerine göre düzenlenerek yöntem uygulanır.

Örnek veri seti üzerinde ortalama değer kullanılarak düzeltme yapılırsa oluşacak yeni veri seti Tablo 1.1'de gösterilmiştir.

Tablo 1.1. Ortalama değeri kullanıldığında oluşan veri seti

Kutu No	Orjinal Veri	Düzenlenmiş Veri
Kutu 1	5,9,15	10,10,10
Kutu 2	22,22,25	23,23,23
Kutu 3	26,29,35	30,30,30

Örnek veri seti üzerinde medyan değeri kullanılarak düzeltme yapılırsa oluşacak yeni veri seti Tablo 1.2'de gösterilmiştir.

Tablo 1.2. Medyan değeri kullanıldığında oluşan veri seti

Kutu No	Orjinal Veri	Düzenlenmiş Veri
Kutu 1	5,9,15	9,9,9
Kutu 2	22,22,25	22,22,22
Kutu 3	26,29,35	29,29,29

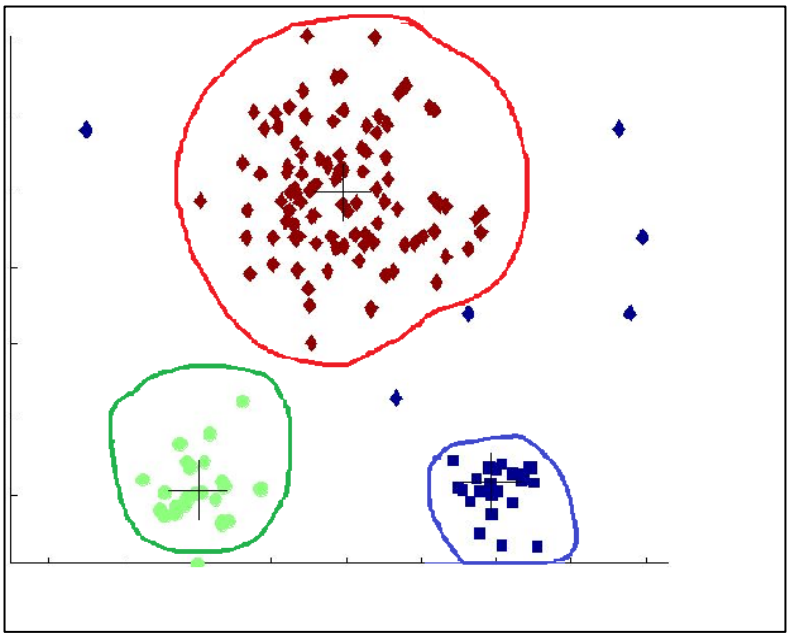
Örnek veri seti üzerinde alt-üst sınır değeri kullanılarak düzeltme yapılırsa oluşacak yeni veri seti Tablo 1.3'te gösterilmiştir.

Tablo 1.3. Alt-üst sınır değeri kullanıldığında oluşan veri seti

Kutu No	Orjinal Veri	Düzenlenmiş Veri
Kutu 1	5,9,15	5,5,15
Kutu 2	22,22,25	22,22,25
Kutu 3	26,29,35	26,26,35

- Kümeleme yöntemi

Aykırı gözlem verilerinin temizlenmesi için kullanılır. Benzer veriler aynı kümede olacak şekilde veriler kümelenir ve aykırı gözlem verileri tespit edilir. Belirlenen veri kümeleri dışında kalan veriler aykırı değer olarak kabul edilir. Tespit edilen aykırı değerler silinerek, veri içerisindeki aykırı veriler temizlenir.



Şekil 1.4. Kümeleme yöntemi örneği

- Regresyon yöntemi

Regresyon yönteminde, veri setine regresyon fonksiyonu uygulanır. En iyi doğru fonksiyonunu bulmayı amaçlar. Bu doğru fonksiyonuna belli bir mesafeden daha uzak gözlemler, aykırı değer olarak kabul edilir. Tespit edilen aykırı değerler silinerek, veri içerisinde tespit edilen gürültülü veriler temizlenir.

1.3.3.2. Veri normalizasyonu

Veri normalizasyonu, farklı ölçeklerde ve çok geniş tanım aralıklarına yayılan verilerin, tek bir düzen içinde ifade edilmesi ve farklı ölçeklerde yer alan veriler üzerinde matematiksel yöntemler uygulayarak, karşılaştırılabilir hale getirilmesidir. Veri normalizasyonu, oluşturulan modelin eğitim sürecini hızlandırabilir (Nayak Vd., 2014). Mesafe ölçümüne dayalı yöntemler için uygulandığında, kullanılan yöntemin başarı oranını artırır (Han vd., 2011). Veri normalizasyonu işlemlerinde kullanılabilecek bazı yöntemler aşağıda açıklanmıştır.

- min-max normalizasyon

Bu yöntemde dönüşüm işlemi uygulanacak veri seti içerisinde en büyük ve en küçük değerler belirlenir. Diğer değerler, tespit edilen en büyük ve en küçük değerler kullanılarak, en büyük değer 1 en küçük değer 0 olacak şekilde normalizasyon işlemi uygulanır. Hesaplamada kullanılan formül denklem (1.1)'de gösterilmiştir. Burada; v normalize edilecek veri, v' normalize edilmiş veri, min veri seti içersindeki en küçük değer, max ise veri seti içersindeki en büyük değeri ifade etmektedir.

$$v' = \frac{v - \min}{\max - \min} \tag{1.1}$$

- Z-Skoru normalizasyon

Veri normalizasyon işlemlerinde yaygın olarak kullanılan bir diğer yöntem Z-skoru normalizasyon yöntemidir. Z-skoru dönüşümde, veri kümesine ait ortalama ve standart sapma değerleri kullanılarak veri normalizasyon işlemi gerçekleştirilir (Tunç ve diğ, 2016). Hesaplamada kullanılan formül denklem (1.2)'de gösterilmiştir. Burada; v

normalize edilecek veri, v' normalize edilmiş veri, μ veri setine ait değerlerin aritmetik ortalaması, σ ise veri setine ait standart sapma değerini ifade etmektedir.

$$v' = \frac{v - \mu}{\sigma} \tag{1.2}$$

1.3.3.2 Veri indirgeme

Hacim olarak büyük boyutlu veri kümeleri üzerinde veri madenciliği yöntemlerinin uygulanması, zaman maliyetini ciddi oranlarda artırmaktadır. Bu durumun önüne geçmek için veri indirgeme teknikleri kullanılır. Veri indirgeme, veri kümesinin daha küçük boyutlu bir örneğini elde etme işlemidir. Hacim olarak daha küçük bir veri kümesi elde edilirken, orijinal veri kümesine ait veri bütünlüğünün korunması önemlidir. Veri indirgeme yöntemleri boyut azaltma, örnek sayısı azaltma ve veri sıkıştırma şeklinde uygulanabilir.

1.3.3.3 Veri bütünleştirme

Farklı kaynaklarda yer alan verilerin, ortak bir veri tipine dönüştürülerek, tek bir veri kaynağında toplanmasıdır. Farklı kaynaklardan toplanan verilerin tutarlı bir veri seti oluşturması için, aynı verilerin tutulduğu özniteliklerin veri tipinin aynı olması gereklidir.

1.3.4. Modelleme

Veri madenciliği sürecinin bu aşamasında, belirlenen problemin çözümüne dönük olarak, veri ön işleme aşamasında düzenlenen veri seti üzerinde, makine öğrenmesi veya istatistiksel bir model geliştirilir. Uygulanan model üzerinde, problemin çözümüne dönük olarak iyileştirmeler gerçekleştirilebilir (Şeker, 2018). Bu aşamada, belirlenen problemin çözümüne dönük olarak, en yakın çıktıları veren modelin belirlenmesi için, birden fazla yöntem uygulanabilir. (Akküçük, 2011).

1.3.5. Değerlendirme

Bu aşamada, sürecin geneli ve uygulanan modelin başarı oranı, başlangıçta belirlenen hedefler açısından incelenir. Değerlendirme aşaması, sonuçların değerlendirilmesi, sürecin değerlendirilmesi ve sonraki adımın belirlenmesi adımlarından oluşur

(Chapman vd., 2000). Bu aşamada, gerçekleştirilen veri madenciliği sürecinin, belirlenen problemin çözümüne uygun olup olmadığı değerlendirilir. Bu adım, elde edilen değerlendirme sonuçlarına göre, sürece ürün aşaması ile mi, yoksa problemin tanımlanması aşaması ile mi devam edileceği kararının verildiği adımıdır (Şeker, 2018). Problemin tanımlanması aşamasında ortaya konan hedeflere istendiği ölçüde ulaşılamadığının tespit edilmesi durumunda, süreç problemin tanımlanması aşamasına dönülerek en baştan tekrarlanır. Elde edilen sonuçların, süreç başlangıcında belirlenen hedefleri karşıladığı tespit edilir ise, sürece uygulama adımı ile devam edilir.

1.3.6. Uygulama

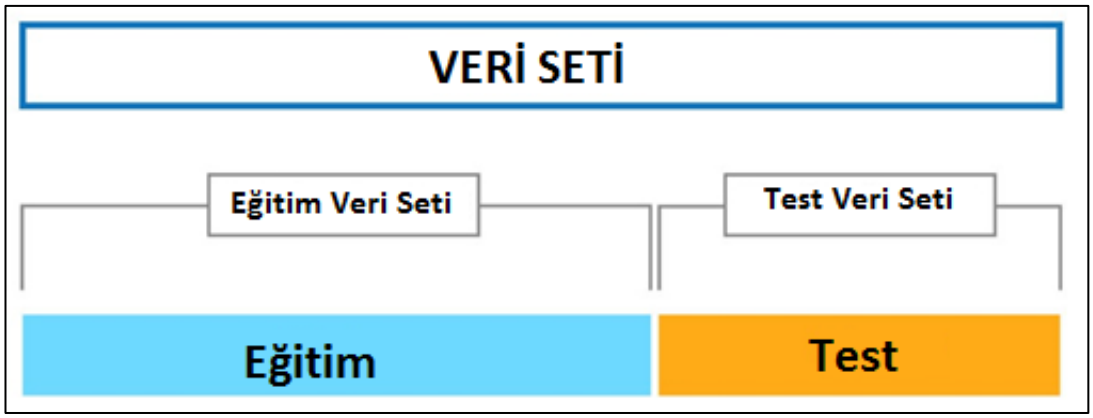
Bu aşamada, veri madenciliği süreci sonucunda oluşan ürün gerçek hayat uygulamalarında kullanıma alınır. Uygulama adımında, sürece dönük izlenme ve bakım çalışmalarına devam edilir. Ortaya çıkan ürün bağımsız olarak kullanılabileceği gibi, farklı sistemler ile bütünleşik olarak da kullanılabilir. Uygulama aşaması, uygulama planının hazırlanması, bakım ve takip sürecinin planlanması, sürecin değerlendirilmesi ve final raporunun hazırlanması alt adımlarından oluşmaktadır (Chapman vd., 2000).

1.4. Model Doğrulama Yöntemleri

Oluşturulan veri madenciliği modelinin eğitilmesi ve değerlendirilmesi için, veri seti, eğitim ve test veri seti olarak ayrılır. Oluşturulan modelin eğitilmesi için eğitim veri seti, test edilmesi için ise test veri seti kullanılır. Bu işlem için veri setinin özelliklerine göre farklı yöntemler kullanılabilir. En sık tercih edilen yöntemler Holdout, K-Katlı Çapraz Doğrulama, Bootstrap ve Leave-One-Out'dur (Bozkır, 2009).

1.4.1. Sınama seti yaklaşımı (Holdout)

Sınama seti yaklaşımı yönteminde veri seti, belli oranlarda test veri seti ve eğitim veri seti olarak iki parçaya ayrılır. Eğitim veri seti kullanılarak model eğitilir, test veri seti kullanılarak oluşturulan modelin performans analizi gerçekleştirilir. Genel olarak veri setinin 2/3'lük kısmı modelin eğitilmesi, kalan 1/3'lük kısmı ise modelin test edilmesi için kullanılır (Keskin, 2018).

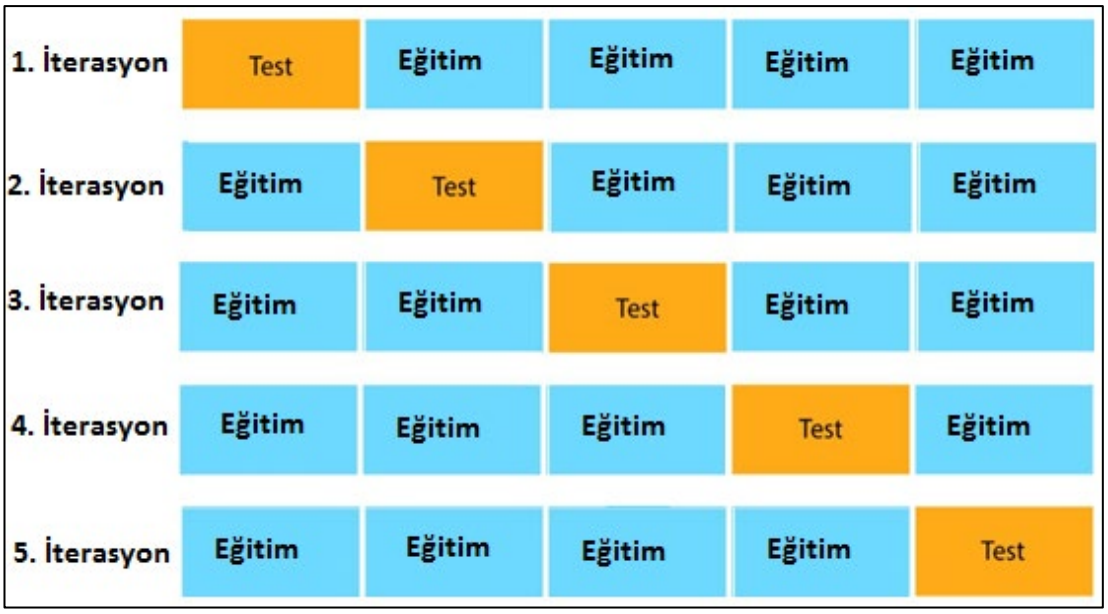


Şekil 1.5. Holdout yöntemi (URL-1)

Sınama seti yaklaşımı yönteminde, örnek sayısının az olması, modelin sağlıklı biçimde test edilememesi problemini ortaya çıkarmaktadır. Bu durumun önüne geçmek için farklı yöntemler geliştirilmiştir (Keskin, 2018).

1.4.2. K-Katlı çapraz doğrulama (K-Fold Cross Validation)

K-katlı çapraz doğrulama yönteminde, veri seti eşit derinlikte k adet parçaya ayrılır. Bir parça test veri kümesi ve kalan parçalar eğitim veri kümesini oluşturur. Daha sonra her parça test veri kümesi ve kalan diğer parçalar eğitim veri kümesi olacak şekilde döngü tekrarlanır. Elde edilen test hatalarının ortalaması, oluşturulan modelin hata oranını verir. Bu yöntemde oluşturulacak grup sayısı (k değeri) kullanıcı tarafından belirlenir (Yakut, 2018).



Şekil 1.6. K-Katlı Çapraz Doğrulama yöntemi (URL-1)

1.4.3. Leave-one-out

K-katlı çapraz doğrulama yönteminden yola çıkılarak geliştirilen "Leave-one-out" yönteminde, veri setindeki örnek sayısı n olarak kabul edilir ise, model n-1 adet örnek ile eğitilir. Dışarda kalan 1 adet örnek, modeli test etmek için kullanılır. Bu işlem, veri setinde yer alan her bir örnek için tekrarlanır. Büyük boyutlu veri setleri için doğru sınıflandırma yapma olasılığını artırır, ancak her örnek için test yapılması zaman maliyetinde ciddi artışlara neden olur (Keskin, 2018).

1.4.4. Yeniden örnekleme yöntemi (Bootstrap)

Bootstrap yöntemi, veri seti içerisinden rastgele veri kümeleri seçilerek, veri kümesinin eğitim ve test kümesi olarak kullanılması yöntemidir. Aynı örnek, farklı veri kümelerinde hem eğitim hem test için kullanılabilir. Elde edilen test hatalarının ortalaması, oluşturulan modelin hata oranını verir.

1.5. Model Başarı Değerlendirmesi

Oluşturulan veri madenciliği modellerinin başarı performanslarının değerlendirilmesi ve en yüksek başarı oranına sahip modelin seçilebilmesi için, Karışıklık Matrisi kullanılmaktadır (Chapman vd., 2000).

Karışıklık Matrisi, oluşturulan modelin performansını analiz etmek için kullanılan, tahmin değerleri ve gerçek değerlerin karşılaştırıldığı bir araçtır. Karışıklık matrisi yapısı, Tablo 1.4'te gösterilmiş ve tablo üzerinde yer alan ifadelere ait açıklamalara aşağıda yer verilmiştir.

Tablo 1.4. Karışıklık Matrisi

	Gerçek Değerler		
Tahmin Değerleri		Pozitif	Negatif
	Pozitif	TP	FP
	Negatif	FN	TN

TP: True Pozitif, test verisindeki değer ile modelin tahmin ettiği sınıf değeri aynı olan örnek sayısını ifade etmektedir. Doğru sınıflandırma yapılmıştır.

TN: True Negatif, test verisindeki değer ile modelin tahmin ettiği sınıf değeri aynı olan örnek sayısını ifade etmektedir. Doğru sınıflandırma yapılmıştır.

FP: False Pozitif, test verisindeki değer ile modelin tahmin ettiği sınıf değeri farklı olan örnek sayısını ifade etmektedir. Yanlış sınıflandırma yapılmıştır.

FN: False Negatif, test verisindeki değer ile modelin tahmin ettiği sınıf değeri farklı olan örnek sayısını ifade etmektedir. Yanlış sınıflandırma yapılmıştır.

Oluşturulan modelin başarı oranı üzerinde daha güvenilir sonuçlar elde etmek için, Karışıklık Matrisindeki sonuç değerleri kullanılarak farklı ölçümler yapılabilmektedir. Bu hesaplama ölçütleri aşağıda açıklanmıştır.

1.5.1. Doğruluk (Accuracy)

Veri madenciliği modelinin çalıştırılması sonucunda doğru olarak sınıflandırılmış veri setinin bütün veri setine oranını ifade etmektedir. Modelin ne oranda doğru tahmin gerçekleştirdiğinin ölçütüdür. Denklem (1.3)'de gösterilen formül ile hesaplanır.

$$\text{Dogruluk} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1.3}$$

1.5.2. Hata oranı (Error Rate)

Veri madenciliği modelinin çalıştırılması sonucunda, hatalı olarak sınıflandırılmış veri setinin, bütün veri setine oranını ifade etmektedir. Modelin ne oranda hatalı tahmin gerçekleştirdiğinin ölçütüdür. Denklem (1.4)'de gösterilen formül ile hesaplanır.

$$\text{HataOranı} = \frac{FP + FN}{TP + FP + TN + FN} \tag{1.4}$$

1.5.3. Kesinlik (Precision)

Veri madenciliği modelinin çalıştırılması sonucunda, doğru olarak sınıflandırılmış pozitif veri setinin, pozitif değere sahip olup pozitif olarak sınıflandırılmış ve negatif değere sahip olup pozitif olarak sınıflandırılmış veri setinin toplamına oranını ifade etmektedir. Model olumlu bir tahmin yaptığında ne sıklıkta doğru olduğunun oranıdır. Denklem (1.5)'de gösterilen formül ile hesaplanır.

$$\text{Kesinlik} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1.5}$$

1.5.4. Hassasiyet (Recall)

Veri madenciliği modelinin çalıştırılması sonucunda, doğru olarak sınıflandırılmış veri setinin, pozitif değere sahip olup pozitif olarak sınıflandırılmış ve pozitif değere sahip olup negatif olarak sınıflandırılmış veri setinin toplamına oranını ifade etmektedir. Modelin pozitif durumları ne oranda başarılı tahmin ettiğinin ölçütüdür. Denklem (1.6)'da gösterilen formül ile hesaplanır.

$$\text{Duyarlılık} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1.6}$$

1.5.5. F-Ölçütü (F-Measure)

Veri madenciliği modeli çalıştırılması sonucunda karışıklık matrisinde yer alan sonuç değerleri kullanılarak hesaplanan kesinlik ve hassasiyet değerlerinin çarpımının 2 katının, kesinlik ve hassasiyet değerlerinin toplamına oranıdır. F-ölçütü kesinlik ve hassasiyet değerlerinin harmonik ortalamasıdır. Denklem (1.7)'de gösterilen formül ile hesaplanır.

$$\text{F} - \text{Ölçütü} = \frac{2 \times \text{Kesinlik.Hassasiyet}}{\text{Kesinlik} + \text{Hassasiyet}} \tag{1.7}$$

2. VERİ MADENCİLİĞİ MODELLERİ VE YÖNTEMLERİ

Veri madenciliği modelleri tanımlayıcı ve tahmin edici modeller olmak üzere iki ana başlık altında incelenebilir (Akpınar, 2000).

Tahmin edici modellerde, sonuçları önceden bilinen verilere dayalı olarak geliştirilen model ile, bilinmeyen sonuçlara sahip veri kümeleri için, sonuç değerlerinin tahmin edilmesi amaçlanmaktadır (Özekes, 2003).

Tanımlayıcı modellerde ise, karar verme sürecinde kullanılabilecek mevcut veri setindeki ilişkilerin keşfedilmesi amaçlanmaktadır (Özekes, 2003).

Veri madenciliği modellerini işlevleri bakımından,

- Sınıflama(classification) ve regresyon (regression)
- Kümeleme(clustering)
- Birliktelik kuralları(sequential patterns)

olmak üzere üç ana guruba ayrılmaktadır. Sınıflama ve Regresyon modelleri tahmin edici, Kümeleme ve Birliktelik kuralları ise tanımlayıcı modellerdir (Akpınar, 2000).

Veri madenciliği modelleri öğrenme biçimleri bakımından, gözetimli öğrenme ve gözetimsiz öğrenme olarak iki grupta incelenebilir.

Gözetimli öğrenme, girdi ve çıktı değerlerini bir arada içeren veri kümeleri için uygulanan bir öğrenme metodudur. Gözetimli öğrenme süreci, eğitim veri seti kullanılarak, kullanılacak veri madenciliği yönteminin eğitilmesi adımı ile başlar. Daha sonra, elde edilen model ve test veri seti kullanılarak tahmin değerleri elde edilir. Tahmin değerleri ve olması gereken değerler karşılaştırılarak, oluşturulan modelin başarı oranı hesaplanır (Bayer vd., 2015).

Gözetimsiz öğrenme ise, sadece girdi değerlerinin yer aldığı veri kümeleri için kullanılır. Kullanılacak veri madenciliği yönteminin eğitim süreci, yalnızca veri setinde yer alan girdi değerlerinin birbirleri ile ilişkileri kullanılarak gerçekleştirilir (Bayer vd., 2015).

2.1. Sınıflandırma ve Regresyon Modelleri

Sınıflandırma, mevcut veri seti üzerinde ileriye dönük veri eğilimlerini keşfetmek için, veri setinden faydalanılarak, veriyi önceden tanımlanmış bir sınıfa dahil etmektir (Han vd., 2011). Kategorik değerlerin tahmininde sınıflandırma, süreklilik gösteren değerlerin tahmininde regresyon modelleri kullanılmaktadır.

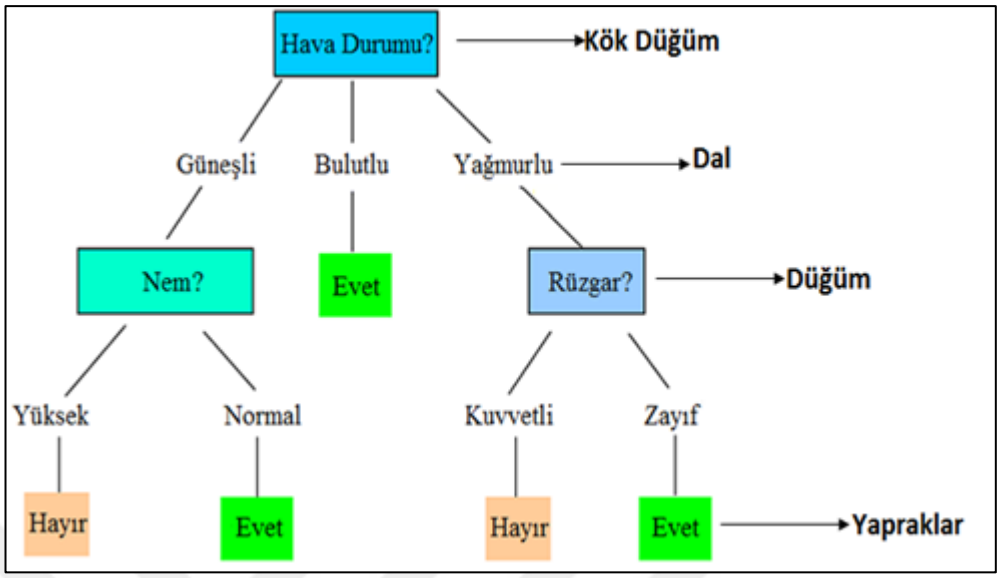
Sınıflandırma modellerinde kullanılan bazı yöntemler;

- Karar Ağaçları (Decision Trees)
- Yapay Sinir Ağları (Artifical Neural Networks)
- Destek Vektör Makineleri (Support Vector Machines)
- Rastgele Orman (Random Forest)
- K-En Yakın Komşu (K-Nearest Neighbor)
- Naive Bayes'dir (Çalık, 2019).

2.1.1. Karar ağaçları

Karar ağaçları, veri madenciliği modellerinde sınıflandırma problemlerinin çözümü için sıklıkla tercih edilen, gözetimli öğrenme yaklaşımlarındandır. Önceden tanımlanmış hedef değerler kullanılarak, yeni gözlemlerin sınıflandırılması amaçlanmaktadır. Düşük maliyetli oluşu, anlaşılmasının ve yorumlanmasının kolaylığı nedeni ile yaygın olarak tercih edilen sınıflandırma yaklaşımlarındandır (Çalış, 2014). Örnek karar ağacı yapısı şekil 2.1'de gösterilmiştir.

Karar ağacı yapısı, düğüm, dal ve yapraklar olmak üzere üç ana kısımdan oluşmaktadır (Kavazoğlu vd., 2010). Ağaç yapısında her düğüm bir özniteliğe karşılık gelmektedir. Karar ağaçlarında ilk düğüme kök düğüm adı verilir. Bir karar ağacı kök düğüm ile başlar ve aşağı doğru ilerleyen dallar ile devam eder. Bazı düğümler uç düğümdür (yapraklar) ve bu düğümlerden sonra başka dal veya düğüm gelmez. Kök düğümden her bir uç düğüme ulaşılabilecek, yalnızca bir yol bulunmaktadır. Yapraklar, denetimli öğrenme yöntemlerindeki tahmin edilmek istenen sınıf değerlerine karşılık gelmektedir (Mather P.M., 2003). Karar ağaçları, hedef değişkenleri (yapraklar) en doğru biçimde tahmin edecek yapıyı oluşturmayı amaçlamaktadır. Örnek karar ağacı yapısı şekil 2.1'de yer almaktadır.



Şekil 2.1. Karar Ağacı örneği (Dalkılıç vd., 2015)

Karar ağaçlarında amaç, oluşturulan ağacı minimum derinlikte tutmaktır. En az derinliğe sahip ağacı elde etmek için, en fazla yayılıma sahip öz niteliğin, kök düğüm olarak seçilmesi gerekmektedir. Karar ağaçları yapısının oluşturulması sürecinde, ağaçtaki dallanmanın hangi yönteme göre yapılacağı belirlenmelidir. Dallanma yöntemi olarak, bilgi kazancı (imformation gain), kazanç oranı (gain ratio) ve Gini indeksi, kullanılan başlıca yöntemlerdendir (Han vd., 2011).

Karar ağacı yaklaşımında, ID3, C4.5, C5.0, CART ve CHAID kullanılan algoritmalardan bazılarıdır (Çalış, 2014). C4.5 ve C5.0 algoritmaları dallanma yöntemi olarak kazanç oranı, ID3 algoritması ise bilgi kazanımı yönteminin kullanıldığı algoritmalardır (Han vd., 2011).

Bilgi kazanımı yöntemi, Amerikalı matematikçi Claude Elwood Shannon tarafından ortaya atılan, bilgi kuramı teorisi temel alınarak geliştirilmiştir. Bilgi kazanımı yöntemi kullanılarak oluşturulan bir karar ağacı yapısında ilk adım, en yüksek bilgi kazanımına sahip özniteliği tespit etmektir. Bilgi kazanımı değeri en yüksek olan öznitelik, kök düğüm olarak belirlenir ve tüm veriler sınıflandırılıncaya kadar süreç, tüm öznitelikler için tekrarlanır (Kilimci, 2018).

Bilgi kazanımı yönteminde, en yüksek bilgi kazanımına sahip özniteliği tespit etmek için, özniteliklere ait entropi değerleri hesaplanır. Entropi, beklenmeyen durumun ortaya çıkma olasılığını, yani belirsizliği ifade eder. Eğer örneklerin tamamı düzenli dağılmış ise entropi değeri 0, eşit dağılmış ise entropi değeri 1, düzensiz bir dağılım söz konusu ise entropi değeri 0 ila 1 arasında bir değer alır.

Sistemin entropisi, denklem (2.1)'de gösterilen formül kullanılarak hesaplanır. Burada; m olasılık sayısı, Pi, D niteliğine ait i. elemanın olasılık değeridir (Yakut, 2018).

$$\mathrm{Info(D)} = -\sum_{i=1}^{m} p\mathrm{i}\mathrm{Log}_2(p\mathrm{i}) \tag{2.1}$$

Daha sonra her bir özniteliğe ait bilgi (imformation) değeri hesaplanır. Bilgi değeri hesaplama formülü Denklem (2.2)'de gösterilmiştir. Burada; D toplam değer sayısını, Dj ise özelliğin aldığı değerlerin sayısını ifade etmektedir. $Info(Dj)$ her bir özellik için hesaplanan entropi değeridir.

$$\mathrm{Info}_A(\mathrm{D}) = \sum_{J=1}^{n} \frac{\mathrm{Dj}}{\mathrm{D}} \mathrm{XInfo(Dj)} \tag{2.2}$$

Son olarak sistemin entropi değeri ve her bir özniteliğe ait bilgi değerlerinin farkı alınarak, her bir özniteliğe ait bigi kazanımı hesaplanır. Bu işlem veri kümesinde yer alan tüm öznitelikler için gerçekleştirilir. A niteliğine ait bilgi kazanımı hesaplama formülü Denklem (2.3)'de gösterilmiştir.

$$\mathrm{Gain(A)} = \mathrm{Info(D)} - \mathrm{Info}_A(\mathrm{D}) \tag{2.3}$$

Tüm özniteliklerin bilgi kazanımı hesaplandıktan sonra, bilgi kazanım değeri en yüksek öznitelik, kök düğüm olarak belirlenir. Hesaplama işlemi diğer öznitelikler için tekrarlanarak ağaç yapısı oluşturulur.

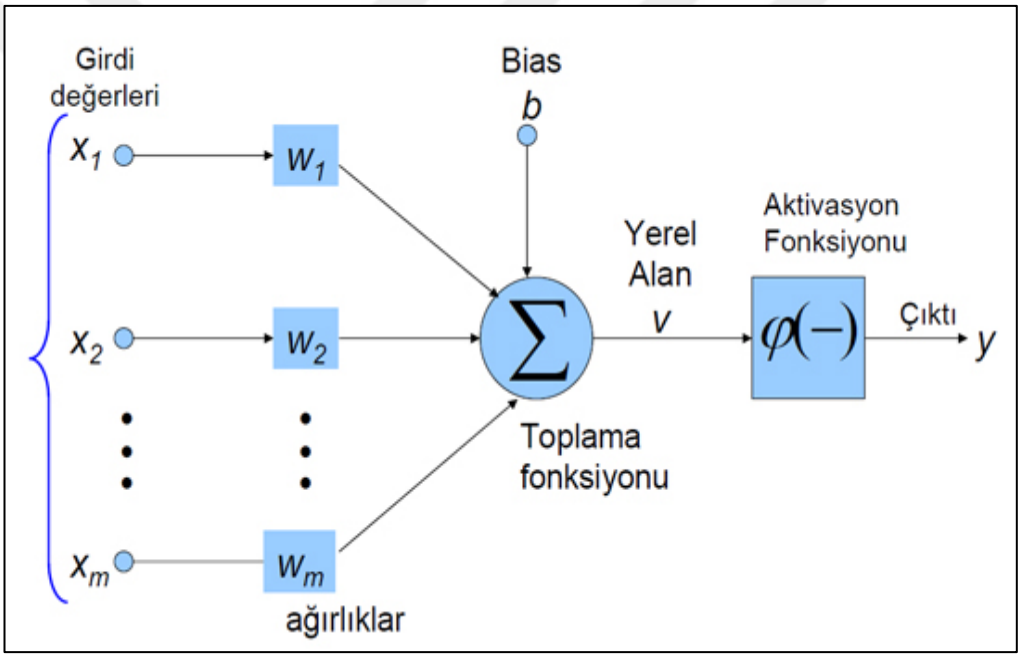
2.1.2. Yapay sinir ağları

Yapay sinir ağları, insan beyninde bulunan sinir hücrelerinin işlev ve çalışma prensiblerinin, matematiksel olarak modellenmesini referans alan, sınıflandırma ve regresyon problemlerinin çözümü için kullanılan, güçlü veri madenciliği

yöntemlerinden bir tanesidir (Akı, 2017). Gerçek bir sinir hücresinin yapısı taklit edilerek oluşturulan yapılardır. Yapay sinir ağları girdi, ağırlıklar, toplam fonksiyonu, aktivasyon fonksiyonu ve çıktılardan oluşur (Terzi, 2009). Örnek bir sinir ağ modeli şekil 2.2'de gösterilmiştir.

Girdi, yapay sinir ağına dış dünyadan, başka hücrelerden veya yapay sinir hücresinin kendisinden gelen veridir.

Ağırlıklar, girdi değerlerinin sistem üzerindeki etkisini ifade eden değerdir. Geldikleri bağlantıların ağırlıklarıyla çarpılarak hesaplanır. Ağırlık değerleri, girdi değerlerine ait önem derecelerini ifade etmez.



Şekil 2.2. Nöron modeli (URL-4)

Girdi değerleri ağırlıklandırıldıktan sonra, toplama fonksiyonuna gönderilir. Toplama fonksiyonu, ağırlıklandırılan girdileri toplayarak, girdi hücrelerinin net girdisini hesaplayan fonksiyondur. Ağırlıklı toplama, minimum, maksimum, çarpım ve kümülatif toplam kullanılan başlıca toplama fonksiyonlarıdır. Her hücrede aynı toplama fonksiyonu kullanılabileceği gibi, farklı hücrelerde farklı toplama fonsiyonları kullanılabilir (Nasuhoğlu, 2019).

Aktivasyon fonksiyonu, net girdi değerini kullanarak, hücre çıktısını belirleyen fonksiyondur. Genelde aktivasyon fonksiyonu olarak sigmoid fonsiyonu kullanılır

(Terzi, 2009). Sigmoid fonksiyonu yanısıra step, tanh, softmax and ReLU sıklıkla kullanılan diğer aktivasyon fonksiyonlarıdır. Sigmoid fonksiyonu denklem (2.4)'de gösterilmiştir.

$$S(x) = \frac{1}{1+e^{-x}} \tag{2.4}$$

Çıktı, aktivasyon fonksiyonu uygulanarak üretilen değerdir. Her hücre birden fazla giriş değerine sahip olmasına ragmen, yalnızca tek çıkış değerine sahiptir. Çıktı değeri dış dünyaya, başka hücrelere veya çıktı değerini üreten hücrenin kendisine girdi değeri olarak gönderilebilir.

Yapay sinir ağları ağ modellerine göre, ileri beslemeli ağlar ve geri beslemeli ağlar olarak iki gurupta incelenebilir (Nasuhoğlu, 2019).

İleri beslemeli ağlar, yalnızca ileriye doğru hareketin söz konusu olduğu ağlardır. Bu yapıda veri, yalnızca kendinden sonra gelen katmana iletilir. Bu modelde yalnızca, bir önceki katmandan gelen değer kullanılır. Bu ağ modelinde, hiyerarşik bir yapı sözkonusudur.

Geri beslemeli ağlar, bir hücrenin çıkışının kendinden önceki katmana veya kendinden sonraki katmana ya da kendi bulunduğu katmana giriş olarak verilebildiği ağ modelleridir.

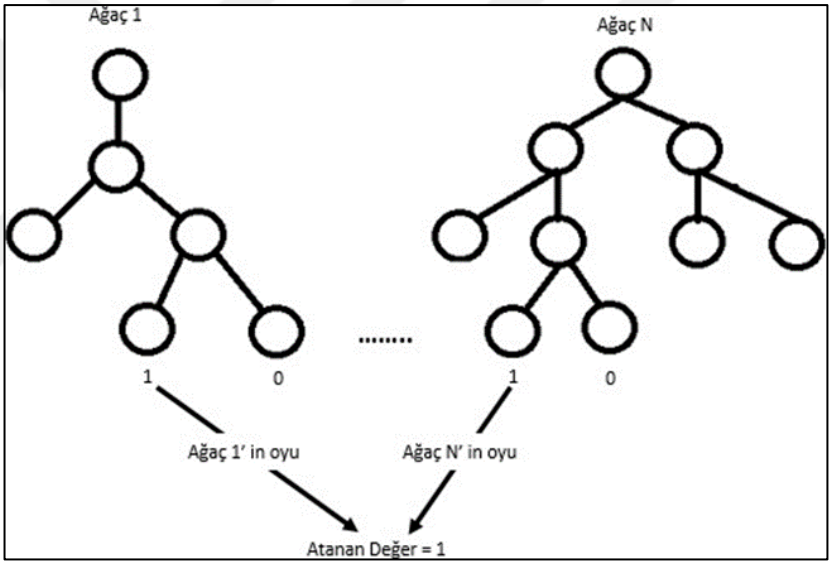
2.1.3. Rastgele orman

Leo Breiman tarafından geliştirilen Rastgele Orman yöntemi denetimli bir sınıflandırma yöntemidir. Sınıflandırma problemlerinin yanı sıra, regresyon problemlerinin çözümü için de sıklıkla tercih edilir. Karar ağaçları yaklaşımından yola çıkılarak geliştirilmiş bir yöntemdir. Karar ağaçları yönteminin en büyük problemlerinden bir tanesi aşırı öğrenmedir. Rastgele orman yöntemi bu problemi çözmek için, rassal olarak belirlenen sayıda karar ağacı oluşturur. Eğitim veri setinden, birden fazla karar ağacı yapısı oluşturarak, sınıflandırma probleminin başarı oranının artırılması amaçlanmaktadır (Fırat vd., 2018, Breiman, 2001). Oluşturulan n adet karar ağacının çoğunluk oyuna göre, sınıflandırılmak istenen örneğin sınıf değeri belirlenir. Birden çok karar ağacı tarafından üretilen tahminler bir araya getirilerek, oluşturulan

karar ağaçlarının çoğunluk oyuna göre, sınıflandırılmak istenen verinin sınıf değerine karar verilmesi mantığına dayanır (Ekelik vd., 2019).

Rastgele orman yöntemi Bagging (Breiman, 2001) ve random subspace (Ho, 1998) yöntemlerinin birleşmesinden oluşmuştur. Ağaçlar için gözlemler bootstrap rastgele örnek seçim yöntemi ile, değişkenler ise subspace yöntemi ile belirlenir.

Rastgele orman yönteminde belirlenmesi gereken parametreler, ormanda kullanılacak ağaç sayısı ve her bir düğüm için kullanılacak değişken sayısıdır. Sınıflandırılmak istenen veriye ait sınıf değeri belirlenirken, her bir ağaca ait hesaplanan hata oranları dikkate alınarak ağaçlara farklı ağırlıklar verilebilir. Örnek rastgele orman yapısı şekil 2.3'de gösterilmiştir.



Şekil 2.3. Rastgele Orman yöntemi örneği

Ağaç sayısının artışı yöntemin başarı oranını artırır fakat bu durum zaman maliyetinin artışına neden olur.

2.1.4. K-En yakın komşu

K-En Yakın Komşu yöntemi, sınıflandırma problemlerinin çözümü için kullanılan, denetimli bir öğrenme yöntemidir. Noktalar arasındaki mesafe ölçümü yapılarak, sınıflandırılmak istenen verinin sınıf değerine karar verilir. K değeri mesafe ölçümünün yapılacağı nokta sayını belirtir ve kullanıcı tarafından belirlenir. K-En yakın komşu yönteminde, yeni eklenen örneğe, eğitim veri seti içindeki en yakın k

adet veri seçilir. Yeni örneğin dahil olacağı sınıf bilgisi, yeni örneğe en yakın k adet komşunun çoğunluk oyuna göre belirlenir (Han vd., 2011). Büyük eğitim kümeleri üzerinde uygulandığında oldukça başarılı sonuçlar elde edilmektedir. Ancak büyük veri setlerinde, örnekler arası mesafenin ölçülmesi yüksek hesaplama maliyetlerine sebep olmaktadır (Uzun, 2007).

K-En Yakın Komşu yaklaşımının performansı için en belirleyici etken, örnekler arası uzaklığın hesaplanmasında kullanılacak olan yöntemdir. Uzaklık hesaplamaları için Öklid (2.5), Manhattan (2.6) ve Minkowski (2.7) sıklıkla kullanılan mesafe hesaplama fonksiyonlarıdır (Taşçı, 2016). Burada; X_i ve Y_i değerleri gözlemlerin konumunu ifade etmektedir.

$$\sqrt{\sum_{i=1}^{k}(X_i - Y_i)^2} \tag{2.5}$$

$$\sum_{i=1}^{k}|X_i - Y_i| \tag{2.6}$$

$$\left[\sum_{i=1}^{k}\left(|X_i - Y_i|\right)^q\right]^{1/q} \tag{2.7}$$

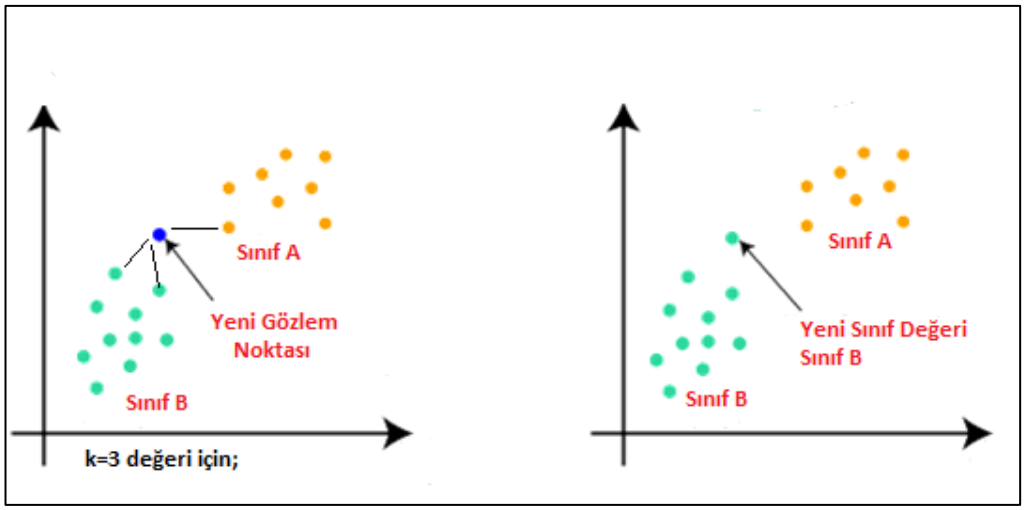
K-En Yakın Komşu yöntemi adımları:

Adım 1: k parametresi belirlenir.

Adım 2: Yeni eklenen verinin, seçilen uzaklık fonsiyonu kullanılarak sınıf etiketi belirlenmiş mevcut verilere uzaklığı hesaplanır.

Adım 3: Hesaplanan uzaklık değerlerine göre yeni eklenen veriye en yakın k adet komşu tespit edilir.

Adım 4: Yeni eklenen verinin sınıf etiketi k adet komşunun çoğunluk oyuna göre belirlenir (Taşçı, 2016).



Şekil 2.4. K-En Yakın Komşuluk örneği

2.1.5. Naive Bayes

Naive Bayes yöntemi, İngiliz matematikçi Thomes Bayes'in ortaya attığı bayes teoremine dayanmaktadır. Bir sınıflandırma probleminin, olasılık hesapları ile çözülebileceği varsayımına dayanmaktadır (Karakoyun vd., 2014). Bayes teoremi denklem (2.8)'de gösterilmiştir. Naive Bayes yönteminde, sınıf etiketi bilinmeyen bir örneğin sınıf etiketinin belirlenmesi için, bayes teoremi kullanılarak her sınıf için, örneğin sınıflara ait olma olasılığı hesaplanır. Hesaplamalar sonucunda olasılık değeri en yüksek sınıf, örneğin sınıf etiketi olarak kabul edilir.

Naive Bayes yönteminde eğitim veri setinin zenginliği, yöntemin doğru sınıflandırma ihtimalini artırmaktadır. Eğitim veri setinin ilk hesaplama maliyetinin yüksek olmasına rağmen, yöntem eğitim aşaması sonrasında hızlı sonuç üretmektedir (Gürmen, 2020). Naive Bayes yönteminde, gözlemlerin önem derecelerinin birbirine eşit olduğu, gözlemlerin birbirinden bağımsız olduğu ve gözlemlerin diğer gözlemler ile ilgili bilgi içermediği kabul edilir.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{2.8}$$

P(A|B) = B olayı meydana geldiğinde A olayının olma olasılığı

P(A) =A olayının olma olasılığı

P(B|A) =*A* olayı meydana geldiğinde B olayının olma olasılığı

P (B) = *B olayının* olma olasılığı

2.2. Kümeleme

Kümeleme yöntemi, tanımlayıcı analitik içerisinde denetimsiz bir yöntemdir. Kümeleme yönteminde, veri kümesi içerisinde yer alan örnekler, benzer özelliklerine ve mesafelerine göre alt kümeler halinde gruplandırılır. Veri setini bölümleme ve alt kümelere ayırma işlemidir. Oluşturulan kümelerin kendi içerisinde homojen, birbirlerine göre heterojen yapıda olması, yani aynı kümede yer alan verilerin benzer özelliklerinin daha fazla, farklı kümelerde yer alan verilerin benzer özelliklerinin daha az olması amaçlanır. Sınıflandırma yönteminden farklı olarak, kümeleme yönteminde önceden belirlenmiş sınıf değerleri yoktur (Göral, 2007).

Kümeleme yöntemlerinin seçiminde, veri setinin özellikleri ve kümeleme yönteminin uygulanış amacı belirleyicidir (Özkes, 2003). Farklı kümeleme yöntemleri, aynı veri seti üzerinde farklı kümeler oluşturabilir. Kümeleme yöntemleri genel olarak 5 ana başlık altında incelenebilir. Bunlar,

1 - Bölme yöntemleri (Partitioning methods)
2- Hiyerarşik yöntemler (Hierarchical methods)
3- Yoğunluk tabanlı yöntemler (Density-based methods)
4- Izgara tabanlı yöntemler (Grid-based methods)
5- Model tabanlı yöntemler'dir. (Model-based methods)

2.3. Birliktelik Kuralları

Birliktelik kuralları, bir arada gerçekleşen olayların analiz edilerek modellenmesi neticesinde elde edilen kurallardır. Bu modeller, geçmiş dönem verilerini kullanarak, olayların birlikte gerçekleşme durumlarını belirler ve bu durumlar üzerinden olasılıksal çıkarımlar yaparlar (Doğan, 2015, Varol Altay vd., 2020). Tanımlayıcı bir yöntemdir. En çok kullanılan birliktelik kuralı yöntemi market sepet analizidir (Bilgiç, 2019).

Birliktelik analizi, tavsiye sistemleri, promosyonal ürün birleştirme, müşteri ilişkileri yönetimi, çapraz satış, web sitesi kullanım madenciliği gibi alanlarda sıklıkla kullanılmaktadır. Birliktelik kuralı analizinde apriori, carma, FP-Growth kullanılan başlıca algoritmalardır.

3. UYGULAMA

Tez calışmasında, bir devlet üniversitesinde eğitim görmekte olan Tıp Fakültesi öğrencilerine ait kurul sınav sonuç verileri, veri madenciliği sınıflandırma yöntemleri kullanılarak incelenmiş ve kullanılan veri madenciliği sınıflandırma yöntemlerine ait başarı oranları karşılaştırılmıştır. Öğrencilere ait dönem sonu başarı durumlarının erken dönemlerde tahmin edilmesi amaçlanmıştır. Ayrıca, eksik veri probleminin çözümüne dönük çalışmalarda, tez çalışmasında kullanılan veri analiz platformunda yer alan yöntemlere ek olarak, k-en yakın komşu veri madenciliği algoritması kullanılmıştır. Eksik veri probleminin çözümüne dönük olarak kullanılan makine öğrenmesi algoritması ve kullanılan veri analiz platformu üzerinde yer alan eksik veri tamamlama yöntemlerinin, oluşturulan veri madenciliği modellerinin başarı oranlarına etkisi araştırılmıştır. Uygulama, CRISP-DM veri madenciliği süreci adımları takip edilerek gerçekleştirilmiştir.

3.1. Uygulama Geliştirme Ortamı

Bir devlet üniversitesinde aktif olarak kullanılmakta olan öğrenci bilgi sisteminden elde edilen veriler, MSSQL kullanılarak oluşturulan veri tabanında toplanmış ve uygulamada kullanılacak verilerin depolandığı veri kaynağı oluşturulmuştur. Eksik veri probleminin çözümüne dönük olarak oluşturulan KNIME düğümleri, Python yazılım dili kullanılarak geliştirilmiştir. Veri madenciliği modelinin oluşturulması, oluşturulan modelin test edilmesi, veri görselleştirme ve veri setine dönük analiz işlemleri, KNIME veri analiz platformu üzerinde gerçekleştirilmiştir.

3.1.1. Python

Python, Hollandalı bir yazılım geliştirici olan Guido van Rossum tarafından, 1990 yılında tasarlanmıştır. Nesne yönelimli, modüler ve etkileşimli, yüksek seviyeli bir programlama dilidir. Kolay öğrenilebilmesi, kolay okunması, bakımının kolay olması ve birçok veri bilimi kütüphanesi bulundurması, dilin veri madenciliği projelerinde kullanımını oldukça yaygınlaştırmıştır. Açık kaynak kod lisansına sahip olan ve

ücretsiz yazılım geliştirilmesine imkanı sunan Python programlama dilinin; Windows, Unix/Linux ve MacOS işletim sistemleri üzerinde farklı yöntemlerle çalıştırılması mümkündür. Pandas, Numpy ve SklLearn gibi birçok açık kaynak kodlu veri madenciliği kütüphanesini bünyesinde barındırmaktadır. Bu durum Python yazılım dilini, veri madenciliği uygulamalarında ön plana çıkarmaktadır (URL-6, URL-7).

3.1.2. KNIME

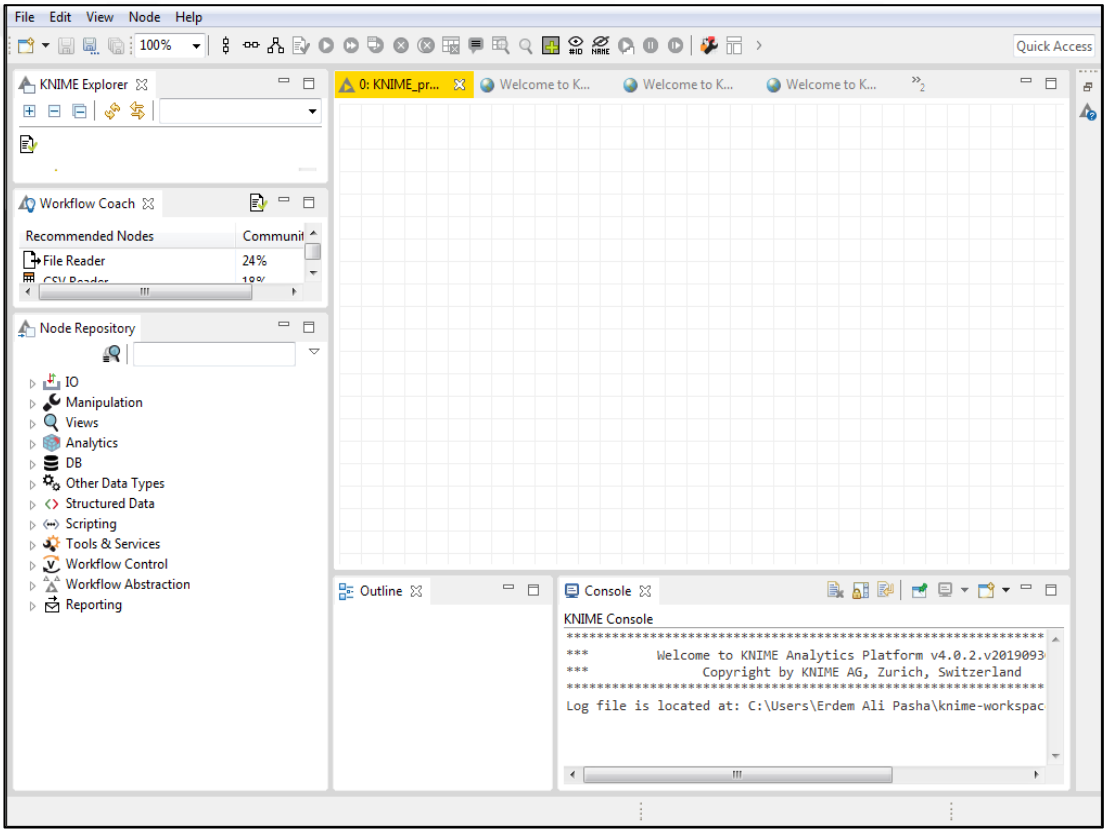
KNIME (Konstanz Information Miner) açık kaynak kodlu veri analiz ve raporlama platformudur. KNIME ocak 2004'te Almanya'da bulunan Konstanz Üniversitesindeki bir yazılım takımı tarafından geliştirilmeye başlandı. İlk KNIME sürümü 2006'da kullanıma alındı ve piyasaya sürülmesinden itibaren değişik sektörlerde faaliyet gösteren birçok farklı şirket ve bireysel kullanıcı tarafından kullanılmaya başlandı (URL-1). KNIME platformu, veri bilimi projelerinin kolay ve hızlı bir biçimde gerçekleştirilmesine olanak sağlamaktadır.

KNIME veri analiz platformu, ağaç tabanlı yöntemler, lojistik regresyon ve derin öğrenme yöntemleri başta olmak üzere, birçok farklı veri analiz yönteminin kullanımına olanak sağlamaktadır. Kullanıcılar, KNIME veri analiz platformu üzerinde görsel olarak iş akışları oluşturabilmektedir. Düğüm adı verilen bileşenler sürükle bırak yöntemi ile çalışır. Düğümler birbirine bağlanarak iş akışları oluşturulur. Gerçekleştirilen projelerin açık kaynak kodlu diğer projeler ile entegrasyon imkanı bulunmaktadır. Oluşturulan projelerde R ve Python gibi yazlım dilleri kullanılarak, kullanıcının ihtiyaçlarına özel düğümler oluşturulabilmektedir. (URL-2)

KNIME projelerinde csv, xls gibi farklı formatlarda saklanan veriler kullanılabileceği gibi, veri tabanları ve veri ambarları ile entegrasyon sağlanarak, oluşturulan projelere veri akışı gerçekleştirilebilir. Ayrıca, KNIME üzerinde oluşturulan veri bilimi projeleri sonucu ortaya çıkan veri kümeleri, oluşturulacak yeni projeler için veri kaynağı olarak kullanılabilir (URL-3). KNIME üzerinde oluşturulan iş akışları sonucu ortaya çıkan veri kümelerinden, doc, ppt, xls, pdf gibi farklı saklama formatlarında raporlar oluşturulabilmektedir (URL-1). KNIME platformu, sahip olduğu modüler yapı sayesinde araştırma ve öğretim özellikleri yanı sıra, ticari projelerde de sıklıkla tercih edilmektedir (Tekerek, 2011). KNIME iş akışı ekranı şekil 3.1'de gösterilmiştir.

Uygulama sürecinde KNIME 3.7.1. versiyonu kullanılmıştır. Modelin eğitildiği ve test edildiği bilgisayara ait işlemci, RAM (Random Acsess Memory) ve işletim sistemi özellikleri aşağıda verilmiştir. Çalışmanın tüm adımlarında aynı bilgisayar ve KNIME sürümü kullanılarak, elde edilen sonuçlar ve sürelerin referans olarak kabul edilmesi amaçlanmıştır.

- İşlemci: Intel Core İ5-7200U
- RAM: 8 GB
- İşletim sistemi: Windows 10
- KNIME version 4.0.2



Şekil 3.1. KNIME iş akış ekranı görüntüsü

3.1.3. MSSQL

Microsoft Structured Query Language (MS-SQL), Microsoft firması tarafından geliştirilen, verilerin belli bir kural ve sistematiğe göre düzenlenerek depolanmasını ve birden fazla kullanıcının erişimine olanak sağlayan, ilişkisel veri tabanı yönetim

sistemidir. MS-SQL server veri tabanı üzerinde işlemler gerçekleştirmek için, T-SQL (Transact-SQL) sorgulama dili kullanılır (Diri, 2017).

3.2. Veri Kümesi

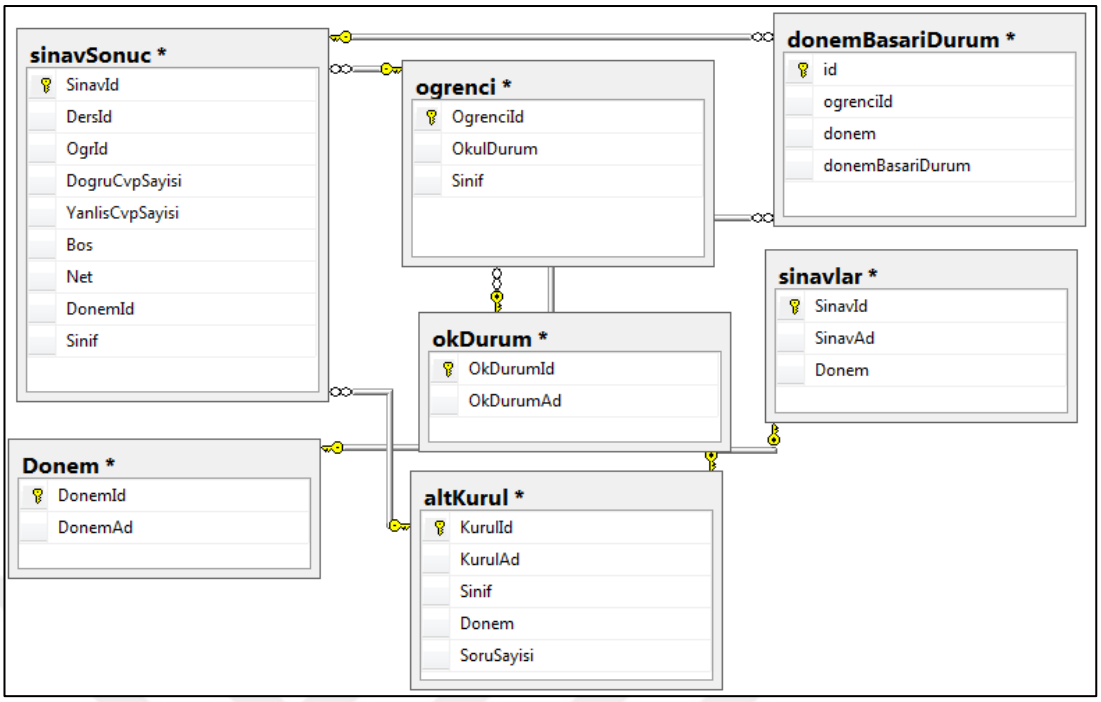
Tıp Fakültesi öğrencilerinin kurul derslerine ait sınav sonuç verileri kullanılarak, başarı durumlarının erken dönemlerde tahmin edilmesi amaçlanan bu tez çalışmasında, bir devlet üniversitesi Tıp Fakültesinde eğtim görmekte olan öğrencilerin, kurul derslerine ait sınav sonuç verileri ve dönem sonu başarı notları kullanılmıştır. Çalışmada, 2016-2017, 2017-2018 ve 2018-2019 eğitim-öğretim yıllarında, dönem 1, dönem 2 ve dönem 3'te eğitim görmüş olan öğrencilere ait veriler kullanılmıştır. Uygulamada kullanılan veri seti, gerekli izinler alındıktan sonra, üniversitede aktif olarak kullanılmakta olan öğrenci bilgi sisteminden temin edilerek, uygulamaya özel olarak oluşturulan veri tabanına kaydedilmiştir. Ayrıca dönem 2 ve dönem 3 öğrencilerinin geçmiş dönemlere ait başarı durumları da veri setine dahil edilmiştir. Veri kümesini oluşturan öğrencilerin sınıf ve yıllara göre sayısal olarak dağılımları Tablo 3.1'de gösterilmiştir.

Tablo 3.1. Veri kümesindeki öğrencilerin sınıf ve yıllara göre dağılımı

Dönem	Sınıf	Öğrenci Sayısı
2016-2017	1	305
2017-2018	2	311
2018-2019	3	304

3.2.1. Veri tabanı tasarımı

Tez çalışmasında kullanılacak verilerin depolanması amacı ile uygulamaya özel veri tabanı oluşturulmuştur. Veri tabanı yönetim sistemi olarak MSSQL kullanılmıştır. Uygulamada elde edilen verilerin direkt olarak oluşturulan veri madenciliği modeline aktarılmayarak, veri tabanı oluşturulmasının temel nedeni, veri standardizasyonunu sağlamaktır. Tez çalışmasında oluşturulan veri tabanı, doğrudan KNIME veri analiz platformuna entegre edilerek, oluşturulan modele düzenli veri akışının sağlanması amaçlanmış ve ileriye dönük çalışmalarda farklı veri tipinde ve saklama formatındaki verilerin, uygulamaya hızlı ve güvenli biçimde enteğre edilmesi için altyapı oluşturulmuştur. Hazırlanan uygulamaya ait veri tabanı ER diyagramı şekil 3.2'de gösterilmiştir.



Şekil 3.2. Veri tabanı ER diyagramı

Oluşturulan veri tabanı, 7 adet tablodan oluşmaktadır. Bunlar "ogrenci", "sinavlar", "altKurul", "sinavSonuc", "okDurum", "donem" ve "donemBasariDurum" tablolarıdır.

3.2.1.1. "Ogrenci" tablosu

"Ogrenci" tablosu, ogrencilere ait verilerin tutulduğu tablodur. Bu tablo içerisinde "ogrenciId", "okulDurum" ve "sinif" alanları yer almaktadır. Bu tablodaki veriler tıp fakültesinde kullanılmakta olan öğrenci bilgi sisteminden temin edilmiştir.

"OgrenciId" alanı, int veri tipinde tanımlanmış ve "Ogrenci" tablosunun birincil anahtarıdır. Orjinal veri seti içersinde OgrenciId bilgisi yer almamaktadır. Uygulama esnasında her öğrenciye uniq bir Id bilgisi atanmıştır.

"OkulDurum" alanında öğrenimine normal süresinde devam eden öğrenciler için True, öğrenimini akademik başarısızlık nedeni ile uzatarak azami sürede bitiremeyecek öğrenciler için False değeri yer almaktadır.

"Sinif" alanında öğrencinin bulunduğu sınıf bilgisi tutulmaktadır ve alan int veri tipinde tanımlanmıştır.

3.2.1.2. “sinavSonuc” tablosu

“sinavSonuc” tablosu öğrencilerin kurul sınavlarına ait sınav sonuç verilerinin yer aldığı tablodur. Bu tablo içerisinde “SinavId”, “DersId”, “OgrenciId”, “DogruCvpSayisi”, “YanlisCvpSayisi”, “Bos”, “Net” alanları yer almaktadır.

“SinavId” alanı int veri tipinde tanımlanmış ve “OgrenciCevaplari” tablosunun birincil anahtarıdır.

”DersId” alanı, “dersler” tablosunun yabancı anahtarı olarak tanımlanmış ve int veri tipindedir.

“OgrenciId” alanı, “ogrenci” tablosunun yabancı anahtarı olarak tanımlanmış ve int veri tipindedir.

“DogruCvpSayisi” alanı öğrencinin ilgili sınavda verdiği doğru cevap sayısı bilgisini tutmaktadır ve int veri tipindedir.

“YanlisCvpSayisi” alanı öğrencinin ilgili sınavda verdiği yanlış cevap sayısı bilgisini tutmaktadır ve int veri tipindedir.

“Bos” alanı öğrencinin ilgili sınavda cevaplamadığı toplam soru sayısı bilgisini tutmaktadır ve int veri tipindedir.

“Net” alanı, öğrencinin ilgili sınavda yanlış cevapladığı toplam soru sayısı 4’e bölünerek toplam doğru cevap sayısından çıkarılması sonucu elde edilen değerdir ve double veri tipindedir.

3.2.1.3. “altKurul” tablosu

“altKurul” tablosu öğrencilerin sınava girdikleri ders kurullarına ait bilgilerin tutulduğu tablodur. Öğrenciler her ders kurulu altında farklı sayıda dersten sınava girmektedir. Bu tablo içerinde “DersId”, “DersAd” ve “Sinif” alanları yer almaktadır.

“DersId” alanı, dersler tablosunun birincil anahtarıdır.

“DersAd” alanında ilgili dersin isim bilgisi tutulmaktadır.

“Sinif” alanında ilgili dersin verildiği sınıf bilgisi tutulmaktadır.

“Donem” alanında dersin verildiği dönem bilgisi yer almaktadır.

“SoruSayisi” alanında ilgili alt kurula ait soru sayısı bilgisi yer almaktadır

4.1.2.4. “dersKurulu” tablosu

“dersKurulu” tablosu öğrencilere uygulanan sınav bilgilerinin tutulduğu tablodur. Bu tablo içersinde “SinavId”, “SinavAd” ve “Donem” alanları bulunmaktadır.

“KurulId” öğrenci tablosunun birincil anahtarıdır ve int veri tipinde tanımlanmıştır.

“KurulAd” alanında sınavlara verilen isimler tutulmaktadır ve varchar tipinde tanımlanmıştır.

“KurulDonem” alanında sınavın yapıldığı dönem bilgisi yer almaktadır.

4.1.2.5. “donemBasariDurum” tablosu

“donemBasariDurum” tablosu, öğrencilerin geçmiş dönemlere ait dönem sonu başarı durumlarının saklandığı tablodur. Bu tabloda yer alan veriler öğrenci bilgi sisteminden alınmıştır. Bu tablo içerisinde “OgrenciId”, “Donem” ve “DonemBasariDurum” alanları tutulmaktadır.

“OgrenciId” alanı öğrenci tablosunun birincil anahtarıdır ve int veri tipinde tanımlanmıştır.

“Sinif” alanında öğrencinin başarı durumunun tutulduğu ilgili sinif bilgisi yer almaktadır.

“DonemBasariDurum” alanında ilgili döneme ait başarı durumu bilgisi tutulmaktadır. Bu alan int veri tipinde tanımlanmıştır.

“BasariNotu” alanında dönemlere ait başarı notu bilgisi tutulmaktadır.

“EgitimYili” alanında öğrencinin başarı durumunun ait olduğu eğitim yılı bilgisi yer almaktadır.

4.1.2.6. “okDurum” tablosu

“okDurum” tablosunda öğrencilerin devam durumları tutulmaktadır. Bu tablo içerisinde “okDurumId” ve “okDurumAd” alanları yer almaktadır. Öğrencilere ait devam durumlarının tutulduğu tablodur.

“okDurumId” alanı “okDurum” tablosunun birincil anahtardır ve int veri tipinde tanımlanmıştır.

“okDurumAd” alanı, öğrencinin okul durum bilgisinin tutulduğu alandır ve varchar tipinde 50 karakter uzunluğunda tanımlanmıştır.

4.1.2.7. “donem” tablosu

“donem” tablosunda kurul sınavlarının uygulandığı dönem bilgileri tutulmaktadır. Bu tablo içerisinde “donemAd” ve “donemId” alanları yer almaktadır.

“donemAd” alanında, ilgili döneme ait isim bilgisi yer almaktadır.

“donemId” alanında, ilgili döneme verilen Id bilgisi yer almaktadır.

3.3. Veri Madenciliği Süreci

Bu tez çalışmasında, CRISP-DM veri madenciliği süreç modeli adımları takip edilerek uygulama geliştirilmiştir. CRISP-DM veri madenciliği süreci, kullanılan platform ve yazılım dilinden tamemen bağımsız olarak, veri madenciliği sürecini standartlaştırmaktadır (Garcia, 2016). Süreç, Problemin Tanımlanması, Verinin Anlaşılması, Veri Ön İşleme, Model Oluşturulması, Değerlendirme ve Uygulama adımlarından oluşmaktadır.

Uygulamaya, problemin tanımlanması adımı ile başlanmıştır. Bu adımda, problemin kaynağı ve problemin çözümüne yönelik beklentiler tespit edilmiştir. Problemin çözümü neticesinde ne tür çıktıların beklendiği ortaya koyulmuştur.

Verinin anlaşılması adımında, çalışmada kullanılacak veri setinin yer aldığı veri kaynakları tespit edilerek, veri toplama işlemi gerçekleştirilmiştir. Yine bu adımda, veri setindeki problemlerin tespitine yönelik olarak, veri analiz işlemleri gerçekleştirilmiştir.

Veri ön işleme adımında, verinin anlaşılması adımında, veri seti içerisinde tespit edilen problemlerin çözümüne yönelik çalışmalar gerçekleştirilmiştir. Tespit edilen eksik veri problemlerine dönük olarak, çözüm yöntemleri belirlenmiş ve uygulanmıştır. Veri seti üzerinde, normalizasyon işlemi gerçekleştrilmiştir.

Modelleme adımı, veri madenciliği yöntemlerinin uygulandığı adımdır. Bu adımda, belirlenen prolemin çözümüne dönük olarak, veri ön işleme adımında düzenlenen veri seti üzerinde, veri madenciliği modelleri oluşturulmuştur.

Değerlendirme adımında, gerçekleştirilen çalışmaların genel bir değerlendirmesi yapılmıştır. Bu adımda, elde edilen değerlendirme sonuçlarının yeterli görülmesi durumunda, sürecin uygulama adımı ile, başlangıçta belirlenen hedeflere uygun olmadığı tespit edilir ise, sürecin problemin tanımlanması adımı ile devam edeceği göz önünde bulundurularak, oluşturulan veri madenciliği modellerinin başarı oranları değerlendirilmiştir.

Uygulama adımı, gerçekleştirilen çalışmanın uyuglamaya alındığı adımdır. Bu adımda veri madenciliği süreci izlenmeye devam eder.

3.3.1. Problemin tanımlanması

Veri madenciliği süreci, çözüm bulunmak istenen problemin net bir biçimde ortaya konulması adımı ile başlar. Problemin doğru ve net bir biçimde ortaya konulması, belirlenen hedeflere tam olarak ulaşılması açısından çok önemlidir (Şeker, 2018).

Kaliteli sağlık hizmeti verilmesinin temel şartı, sağlık çalışanlarının kaliteli bir eğitim almasıdır (Abuhanoğlu vd., 2012). Tıp eğitimi süreçlerinde, büyük hacimli veri kümeleri oluşmaktadır. Oluşan bu veri kümeleri üzerinde, veri madenciliği yöntemleri kullanılarak elde edilen anlamlı bilgi, eğitim-öğretim süreçlerinin planlanmasında, karar süreçlerinde ve öğrencilerin akademik başarılarını artırmak amacı ile kullanılarak, eğitim-öğretim süreçlerinin kalite ve verimliliği artırılabilir mi? Bunun yanısıra, dönem bazlı olarak başarılı ve başarısız öğrenci sayıları erken dönemlerde tahmin edilerek, tıp eğitiminde sahip olunan fiziki mekan ve öğretim üyesi gibi sınırlı kaynaklara yönelik planlamalar daha verimli biçimde yapılabilir mi? problemlerine çözüm aranmıştır. Ayrıca, KNIME veri analiz platformu altında yer alan eksik veri

probleminin çözümünde kullanılan mevcut yöntemlere ek olarak, eksik veri probleminin çözümünde makine öğrenmesi algoritmaları kullanılarak, oluşturulan modellerin başarı oranları artırılabilir mi?

3.3.2. Verinin anlaşılması

Çalışma kapsamında, bir kamu üniversitesi Tıp Fakültesinde, 2016-2017, 2017-2018 ve 2018-2019 eğitim-öğretim yıllarında eğitim görmüş öğrencilerin, eğitimlerinin ilk 3 yılında aldıkları kurul derslerine ait sınav sonuç verileri ve dönem başarı durumlarına ait 22170 satırdan oluşan veri seti kullanılmıştır. Hedef değişken olarak öğrencilerin dönem sonu başarı durumları kullanılmıştır.

Ders kurulu, tıp eğitiminin ilk üç döneminde birbiriyle alakalı sistem ya da konu gruplarından meydana gelen ve ilgi alanındaki bilgi ve becerileri kazandırmayı amaçlayan eğitim süreçleridir. Ders kurulu sınavı, her ders kurulu sonunda, ilgili ders kuruluna ait bilgi ve beceriyi ölçmeyi amaçlayan süreçtir. Ders kurulu teorik sınavları, çoktan seçmeli sınavlar olarak gerçekleştirilmektedir. Farklı zaman aralıklarına yayılan kurul eğitimleri sonunda, ders kurulu altında yer alan alt kurul derslerine ait farklı sayılarda sorunun yer aldığı sınavlar uygulanmaktadır.

Ders kurulu sınavlarına ait sınav sonuç verileri, üniversitede aktif olarak kullanılmakta olan öğrenci bilgi sistemi veri tabanından, gerekli izinler alınarak temin edilmiştir. Çalışmada, 920 öğrenciye ait, 10 ders kurulu altında yer alan, 61 alt kurul dersine ait sınav sonuç verisi ve dönem sonu başarı notları kullanılmıştır. Üniversite öğrenci bilgi sistemi veri tabanından temin edilen veriler, çalışma kapsamında oluşturulan veri tabanına kaydedilmiştir. Oluşturulan veri tabanı, KNIME veri analiz paltformuna bağlanmıştır. Elde edilen verilerin, veri tabanına kaydedilerek, veri düzenleme sürecinin kısaltılması ve oluşturulan modele düzenli veri akışı sağlaması amaçlanmıştır.

2016-2017 döneminde, dönem 1'de eğitim gören öğrencilerin, ilk üç ders kurulu altında yer alan, 21 alt kurul dersine ait, sınav sonuç verileri kullanılarak, öğrencilerin dönem 1 eğitimleri sonundaki, akademik başarı ve başarısızlarının erken dönemde tahmin edilmesi amaçlanmıştır.

2017-2018 döneminde, dönem 2'de eğitim gören öğrencilerin, ilk üç ders kurulu altında yer alan, 14 alt kurul dersine ait sınav sonuç verileri ve dönem 1 eğtimleri sonunda oluşan dönem sonu başarı notları kullanılarak, dönem 2 eğitimleri sonundaki akademik başarı ve başarısızlarının, erken dönemde tahmin edilmesi amaçlanmıştır.

2018-2019 döneminde, dönem 3'de eğitim gören öğrencilerin, ilk 4 ders kurulu altında yer alan, 26 alt kurul dersine ait sınav sonuç verileri, ayrıca dönem 1 ve dönem 2'ye ait dönem sonu başarı notları kullanılarak, dönem 3 eğitimleri sonundaki, akademik başarı ve başarısızlarının erken dönemde tahmin edilmesi amaçlanmıştır.

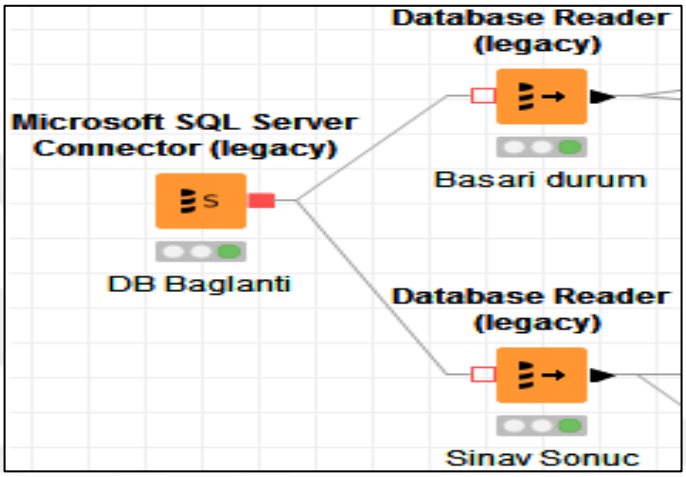
Eksik veri problemine dönük uygulanan yöntemlerin başarı oranlarının değerlendirilmesi amacı ile veri setinin %10'luk kısmı rastgele olarak silinmiştir. Veri seti üzerinde silme işleminin gerçekleştirilmesi için Python Script düğümü kullanılarak Python kodları yazılmıştır. Python Script düğümü özelliklerine veri ön işleme bölümünde ayrıntılı olarak değinilmiştir.

3.3.2.1. Veri setinin uygulamaya dahil edilmesi

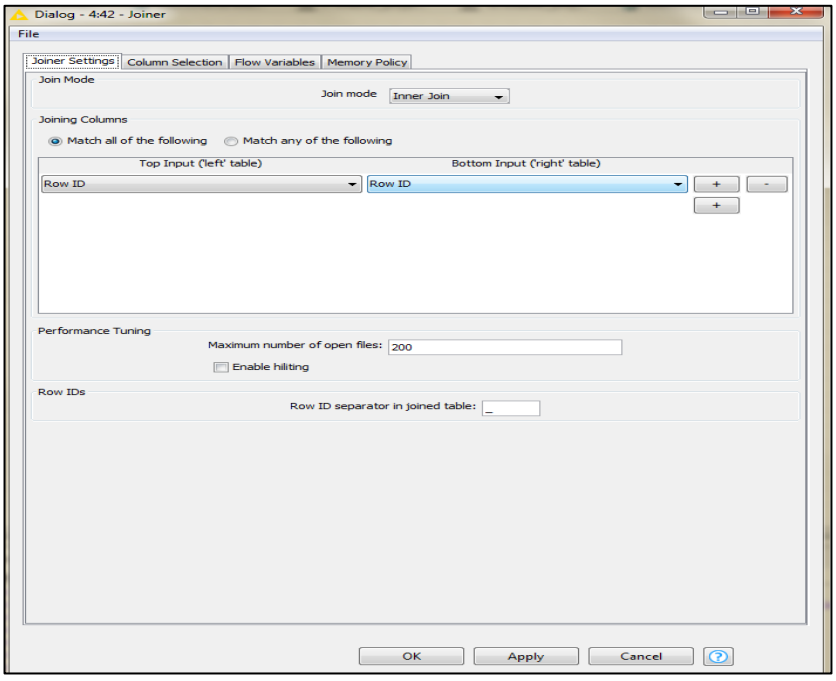
Veri setinin uygulamaya dahil edilmesi sürecinde, KNIME platformunda yer alan Microsoft Sql Server Connector düğümü kullanılmıştır. Bu düğüm kullanılarak oluşturulan veri tabanı ile KNIME platformu bağlantısı sağlanmıştır. Microsoft SQL Server Connector düğümü, Microsoft SQL sunucusuna bağlantı oluşturur. Microsoft SQL sunucusuna bağlantı işleminin gerçekleşmesi için, Microsoft SQL Server Conector düğümü yapılandırma ekranı üzerinden, ana bilgisayar adı, bağlantı noktası ve bağlantı gerçekleştirilmek istenen veri tabanı adının belirtilmesi gerekmektedir. Microsoft SQL Server Conector düğümü yapılandırma ekranı üzerinde gerekli ayarlamalar yapılarak veri tabanı ve KNIME platformu bağlantısı sağlanmıştır. Microsoft SQL Server Connector düğümü yapılandırma ekranı şekil 3.5'de gösterilmiştir.

Database Reader düğümü, Database Reader yapılandırma ekranı kullanılarak oluşturulan sql sorgusunu, Microsoft SQL Server Conector düğümü kullanılarak bağlantı sağlanan veri tabanı üzerinde yürütür ve sonucu bir KNIME veri tablosuna alır. Her dönem için kullanılacak veri seti, sql komutları kullanılarak oluşturulan veri tabanından çekilmiş ve uygulamaya dahil edilmiştir. Database Reader düğümü

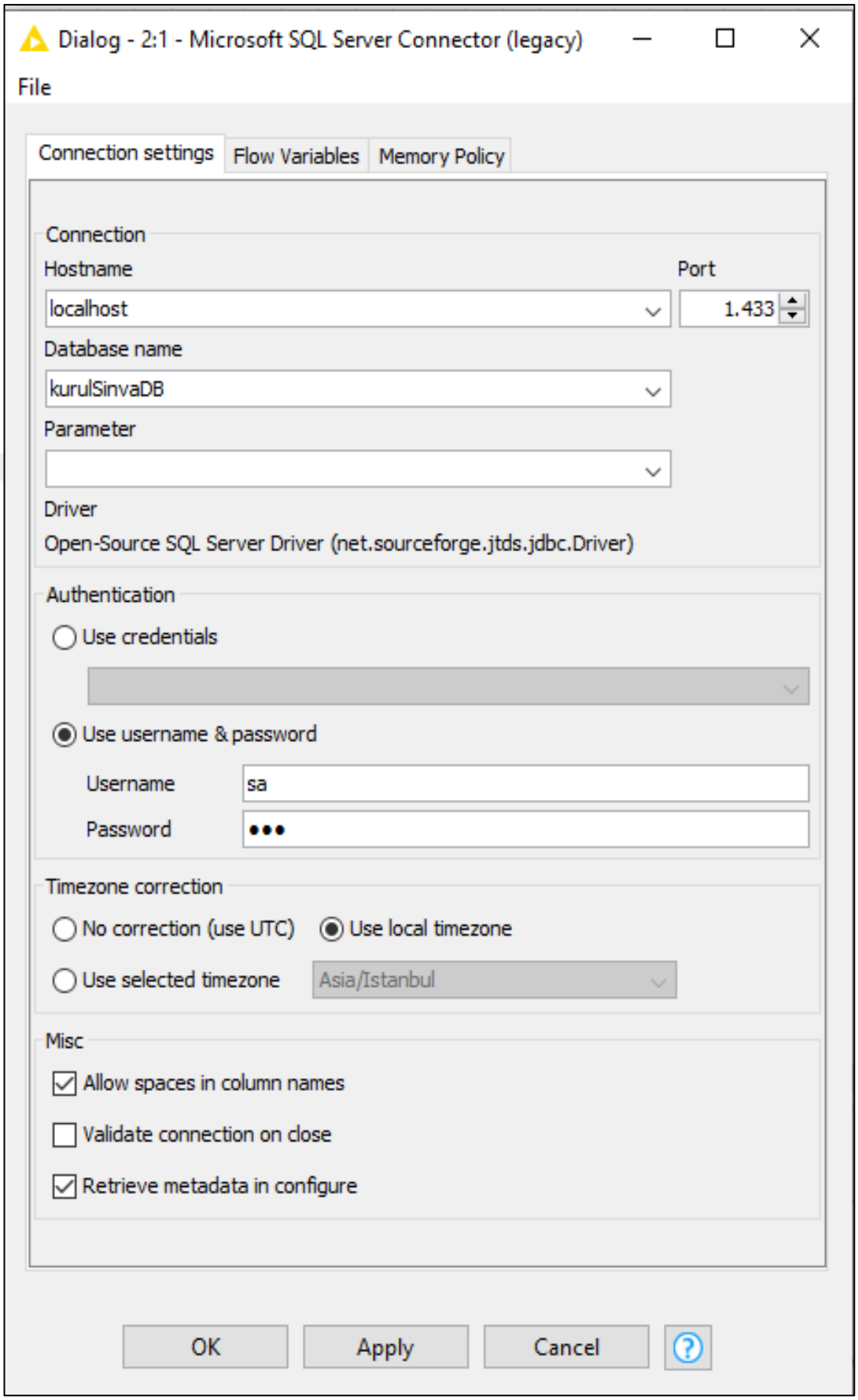
yapılandırma ekranı şekil 3.6'da yer almaktadır. Öğrencilere ait sınav sonuç verileri ve başarı durumlarına ait verilerin uygulamaya dahil edilmesi sürecinde iki ayrı Database Reader düğümü kullanılmış daha sonra elde edilen veriler Joiner düğümü kullanılarak birleştirilmiştir. Joiner düğümü, iki farklı KNIME tablosunu seçilen kolonlar üzerinden veri tabanı benzeri bir yapıda birleştirir. Birleştirme işlemi Row ID alanı üzerinden gerçekleştirilmiştir. Joiner düğümü yapılandırma ekranı Şekil 3.4'de gösterilmiştir. Microsoft SQL Server Conector ve Database Reader düğümleri şekil 3.3'de gösterilmiştir.



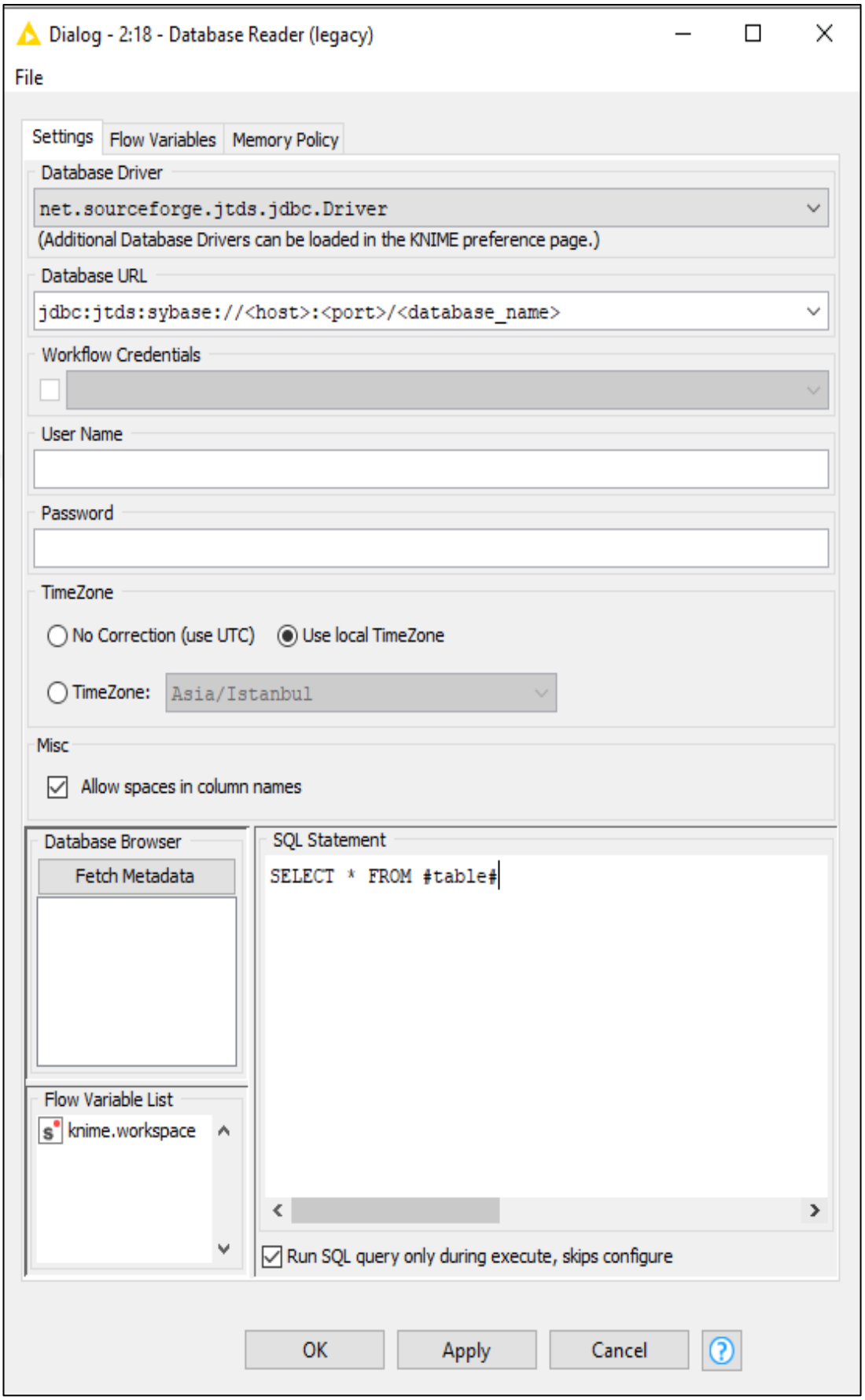
Şekil 3.3. Veri Tabanı Bağlantı ve Sorgu Düğümleri



Şekil 3.4. Joiner düğümü yapılandırma ekranı



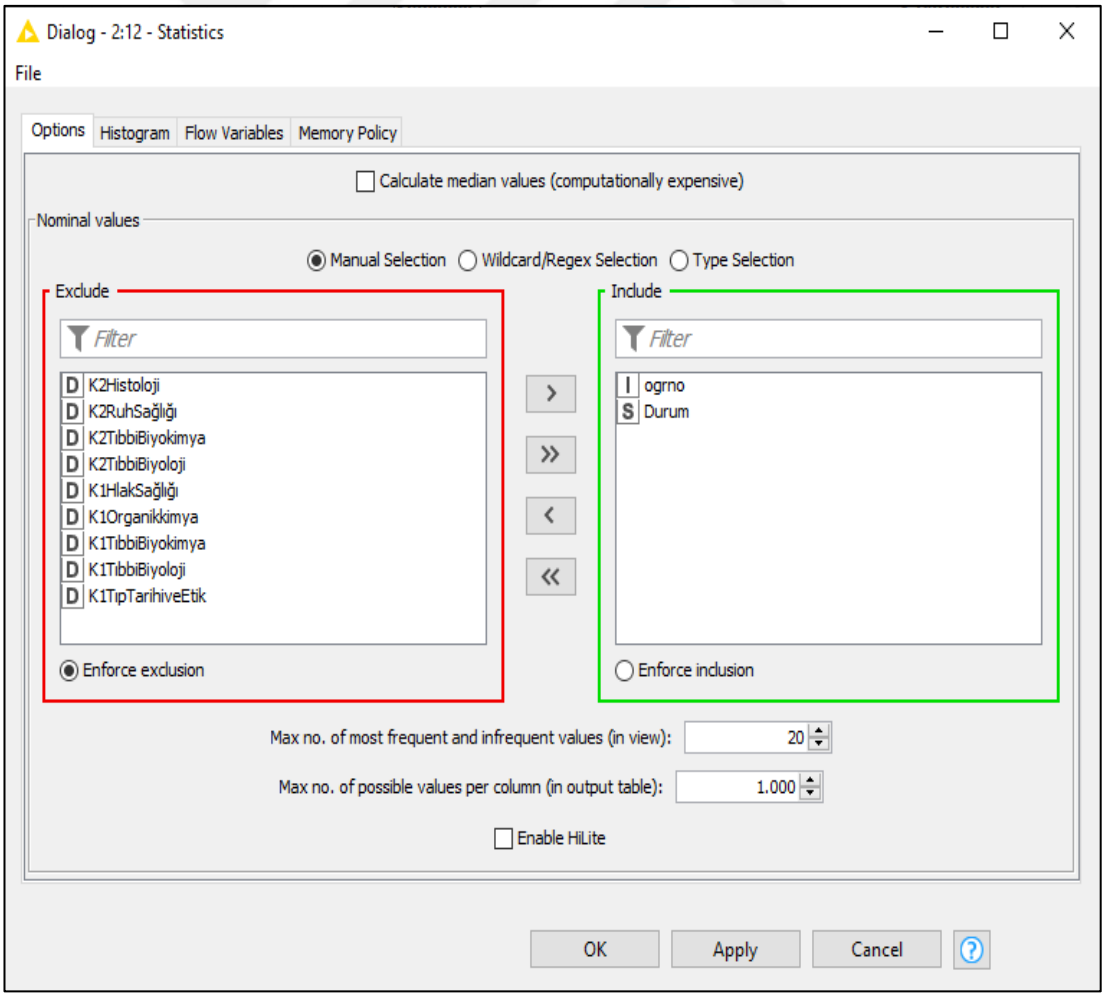
Şekil 3.5. Microsoft SQL Server Connector düğümü yapılandırma ekranı



Şekil 3.6. Database Reader düğümü yapılandırma ekranı

3.3.2.2. Veri setinin analiz edilmesi

Veri setinin anlaşılması ve veri seti içersinde yer alan eksik verilerin tespiti için, KNIME veri analiz platformunda yer alan Statistics düğümü kullanılmıştır. Bu düğüm kullanılarak, veri setine dönük istatistiki bilgiler elde edilmiştir. Statistics düğümü yapılandırma ekranı şekil 3.7'de gösterilmiştir. Bu düğüm, sayısal sütunlardaki minimum, maksimum, standart sapma, varyans, medyan gibi istatistiksel değerlerin yanı sıra, veri seti içerisinde yer alan eksik değerlerin sayısını kolon bazlı olarak verir. Eksik veri probleminin giderilmesine dönük çalışmalarda, Statistics düğümünden elde edilen analiz sonuçları kullanılmıştır. Bu düğüm kullanılarak elde edilen analiz sonuçları, Tablo 3.2'de verilmiştir. Tablo 3.2'de alt kurul derslerine ait ortalama, standart sapma ve eksik değer bilgileri yer almaktadır. Dönemlere ait başarılı-başarısız öğrenci dağılımları ve ders kurulu sınavlarındaki alt kurul derslerine ait soru sayısı dağılımları, şekil 3.8 – 3.11'de yer almaktadır.



Şekil 3.7. Statistics düğümü yapılandırma ekranı

Tablo 3.2. Alt kurul dersleri istatistik bilgileri

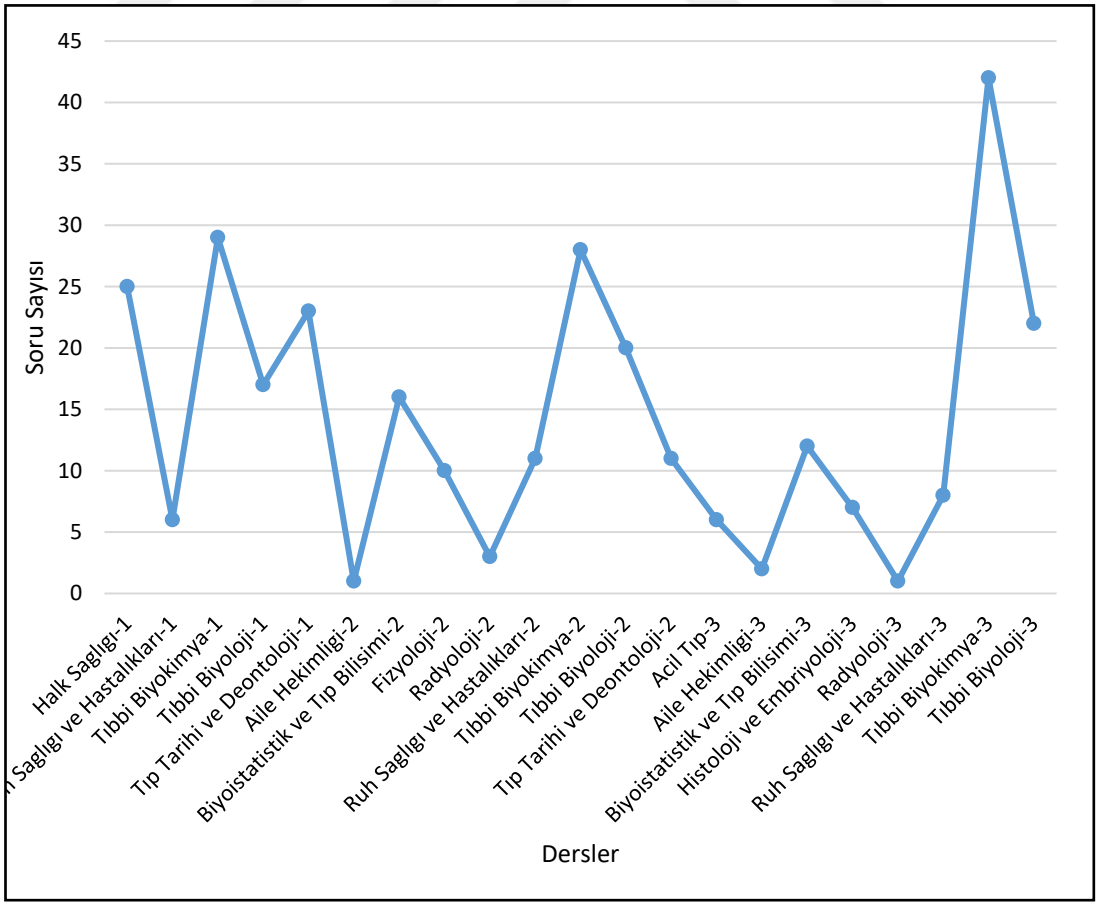
Dönem	Kurul Adı	Alt Kurul Adı	Soru Sayısı	En Düşük	En yüksek	Ortalama	Standart Sapma	Eksi Değer
Dönem 1	Hücre Bilimleri 1	Halk Saglıgı	25	2.75	24	19.13	3.23	29
		Ruh Saglıgı ve Hastalıkları	6	-0.25	6	3.5	1.308	32
		Tıbbi Biyokimya	29	1.5	29	23.08	5.133	30
		Tıbbi Biyoloji	17	0.75	16	10.36	2.563	32
		Tıp Tarihi ve Deontoloji	23	-1.5	21.75	16.99	3.596	31
Dönem 1	Hücre Bilimleri 2	Aile Hekimligi	1	0.25	1	0.808	0.451	32
		Biyoistatistik ve Tıp Bilisimi	16	-0.25	15	10.147	2.825	33
		Fizyoloji	10	-0.75	8	5.399	1.901	30
		Radyoloji	3	-0.75	3	2.718	0.667	31
		Ruh Saglıgı ve Hastalıkları	11	-1.5	11	7.233	2.388	33
		Tıbbi Biyokimya	28	-3	27	15.32	6.301	30
		Tıbbi Biyoloji	20	2.5	20	17.927	2.477	32
Dönem 1	Hücre Bilimleri 3	Acil Tıp	6	-0.25	6	3.676	1.333	32
		Aile Hekimligi	2	-0.5	2	0.095	0.795	33
		Biyoistatistik ve Tıp Bilisimi	12	-1.75	12	6.905	2.98	31
		Histoloji ve Embriyoloji	7	-0.5	7	5.046	1.879	33
		Radyoloji	1	-0.25	1	0.794	0.465	33
		Ruh Saglıgı ve Hastalıkları	8	-2	8	5.459	1.51	29
		Tıbbi Biyokimya	42	-0.5	42	26.926	9.361	32
		Tıbbi Biyoloji	22	-2.75	21	13.947	3.978	32
Dönem 2	Dolaşım ve Solunum Sistemi	Aile Hekimligi	1	-0.25	1	.0582	0.591	29
		Anatomi	18	-3.25	18	9.348	4.493	29
		Fizyoloji	47	13.25	44.5	27.757	6.093	28
		Histoloji ve Embriyoloji	25	1.5	22.75	14.68	4.321	27
		Kardiyoloji	2	-0.5	2	0.377	0.852	28
		Tıbbi Biyokimya	7	-1.75	7	3.351	1.84	29
		Fizyoloji	17	2	17	10.863	3.135	31

Tablo 3.2.(Devam) Alt kurul dersleri istatistik bilgileri

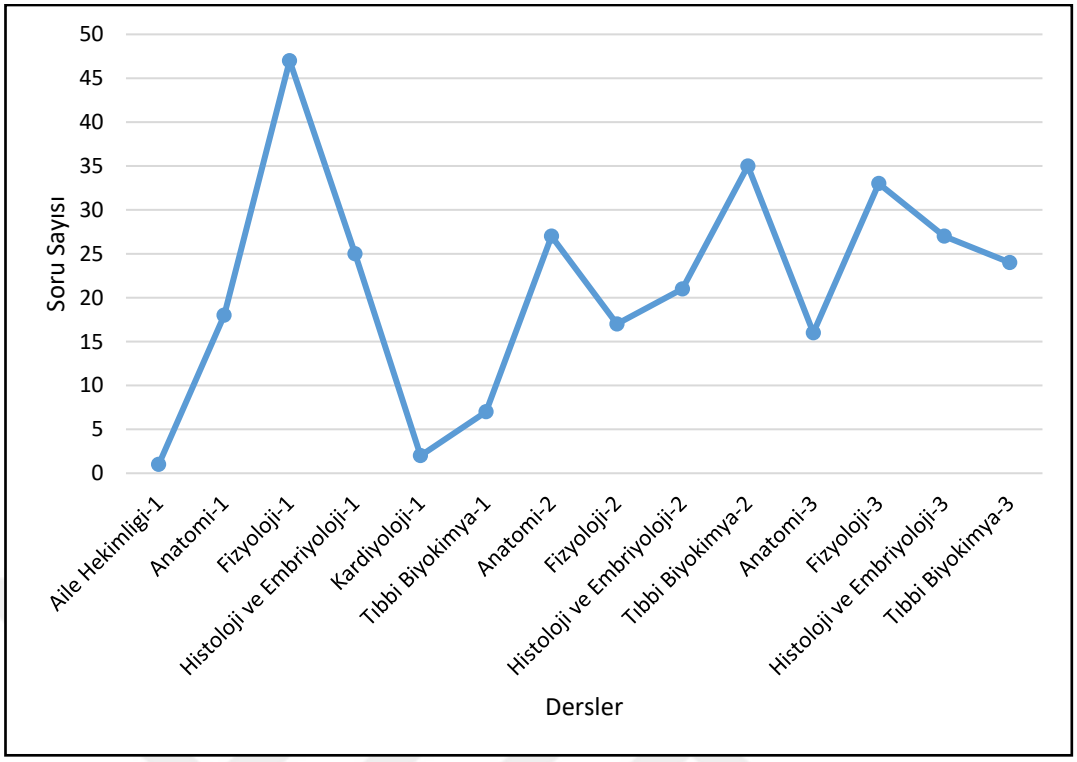
	Ürogenital ve Endokrin Sistem	Histoloji ve Embriyoloji	21	2.25	21	17.02	3.544	27
		Tıbbi Biyokimya	35	5	33.75	22.618	6.424	30
Dönem 2	Sindirim Sistemi ve Metabolizma	Histoloji ve Embriyoloji	27	0.75	27	16.857	5.251	29
		Anatomi	17	-0.25	16	10.299	3.086	28
		Fizyoloji	33	2	33	23.183	6.005	30
		Tıbbi Biyokimya	24	-2.25	22.75	11.409	4.786	29
Dönem 3	Hastalıkların Biyolojik Temelleri	Enfeksiyon Hastalıkları ve Klinik Mikrobiyoloji	6	-0.25	6	3.364	1.431	30
		Radyasyon Onkolojisi	7	-0.5	7	5.553	1.319	27
		Tıbbi Farmakoloji	22	-1.75	22	13.584	5.3	27
		Tıbbi Genetik	11	-1.5	11	8.057	1.855	28
		Tıbbi Mikrobiyoloji	26	-1.25	22.5	14.46	4.82	28
		Tıbbi Patoloji	22	0.75	22	14.712	4.82	28
Dönem 3	Dolaşım ve Solunm Sistemi Hastalıkları	Cocuk Saglıgı ve Hastalıkları	9	-1	9	5.629	2.036	28
		Gögüs Cerrahisi	7	-1.75	7	3.546	1.751	28
		Gögüs Hastalıkları	12	-1.5	11	7.043	2.238	29
		Kardiyoloji	16	-0.25	14.75	10.8	2.396	30
		Tıbbi Farmakoloji	18	-0.75	18	10.84	3.904	29
		Tıbbi Mikrobiyoloji	18	0.5	18	10.946	3.455	29
		Tıbbi Patoloji	12	-0.5	12	8.146	2.899	29
Dönem 3	Sindirim Sistemi ve Hematopoetik Sistem Hastaliklari	Cocuk Saglıgı ve Hastalıkları	13	-1.75	12	6.634	2.704	32
		Enfeksiyon Hastalıkları ve Klinik Mikrobiyoloji	5	0	5	3.632	1.053	30
		Genel Cerrahi	4	0.25	4	3.155	0.949	32
		İç Hastalıkları	20	-2.25	19	10.375	3.885	29

Tablo 3.2.(Devam) Alt kurul dersleri istatistik bilgileri

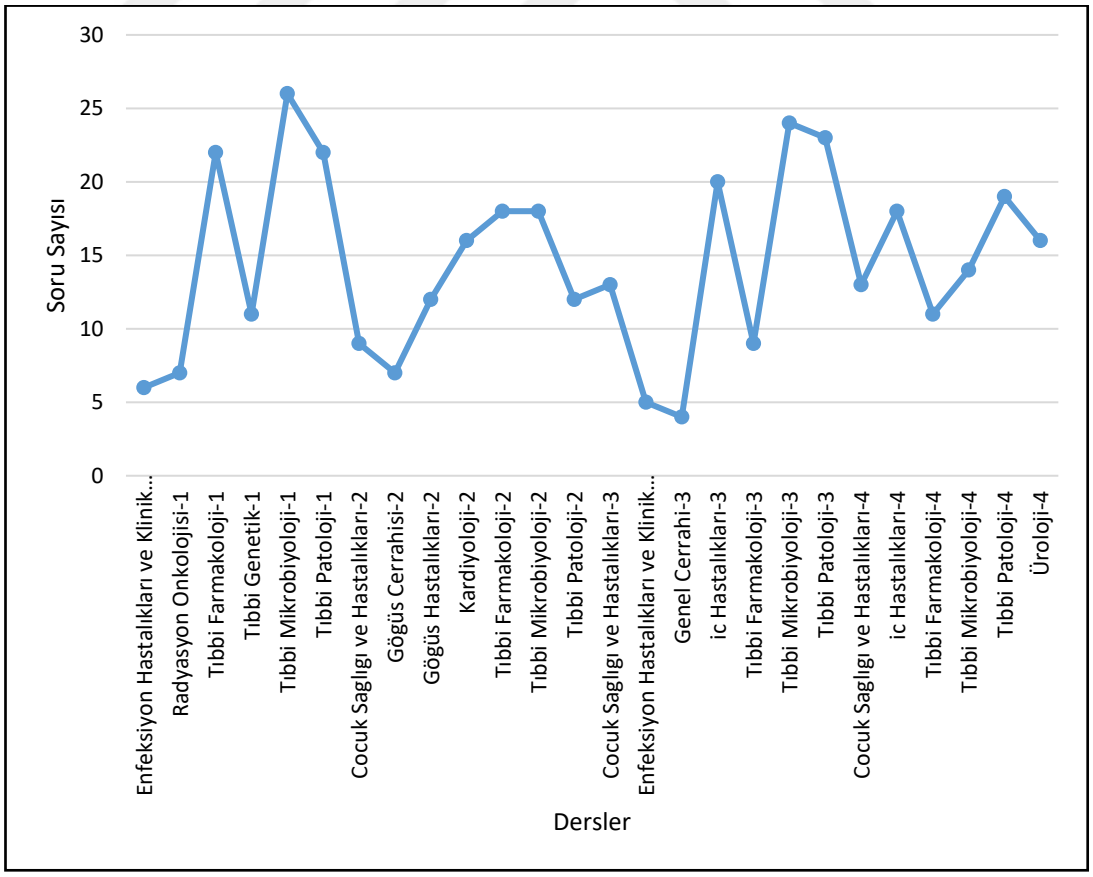
		Tıbbi Farmakoloji	9	-2.25	9	5.067	2.653	30
		Tıbbi Mikrobiyoloji	24	-1.75	22	9.658	2.956	29
		Tıbbi Patoloji	23	-2	21	9.852	3.652	30
Dönem 3	Üriner Sistem	Cocuk Saglıgı ve Hastalıkları	13	-0.75	13	9.005	2.748	30
		ic Hastalıkları	18	0.5	18	11.189	2.285	29
		Tıbbi Farmakoloji	11	-1.5	11	9.044	2.11	30
		Tıbbi Mikrobiyoloji	14	-1	14	8.373	3.843	29
		Tıbbi Patoloji	19	-1	19	10.761	3.323	30
		Üroloji	16	2.75	14.75	8.819	3.5	28



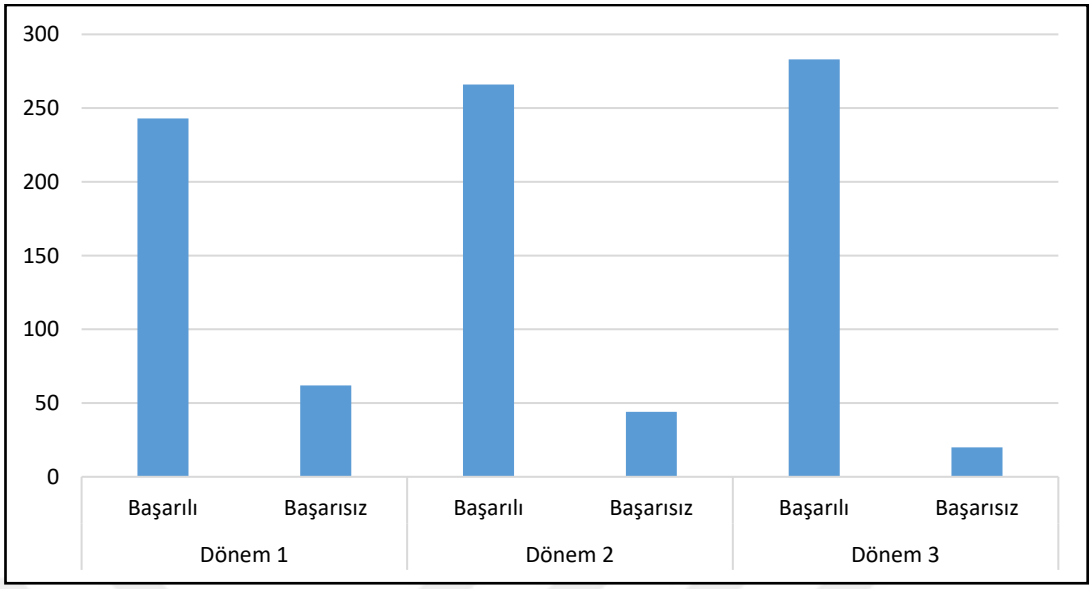
Şekil 3.8. Dönem 1 alt kurul dersleri soru dağılımı



Şekil 3.9. Dönem 2 alt kurul dersleri soru dağılımı



Şekil 3.10. Dönem 3 alt kurul dersleri soru dağılımı



Şekil 3.11. Dönem 1-2-3 başarı durum dağılımları

3.3.3. Veri ön işleme

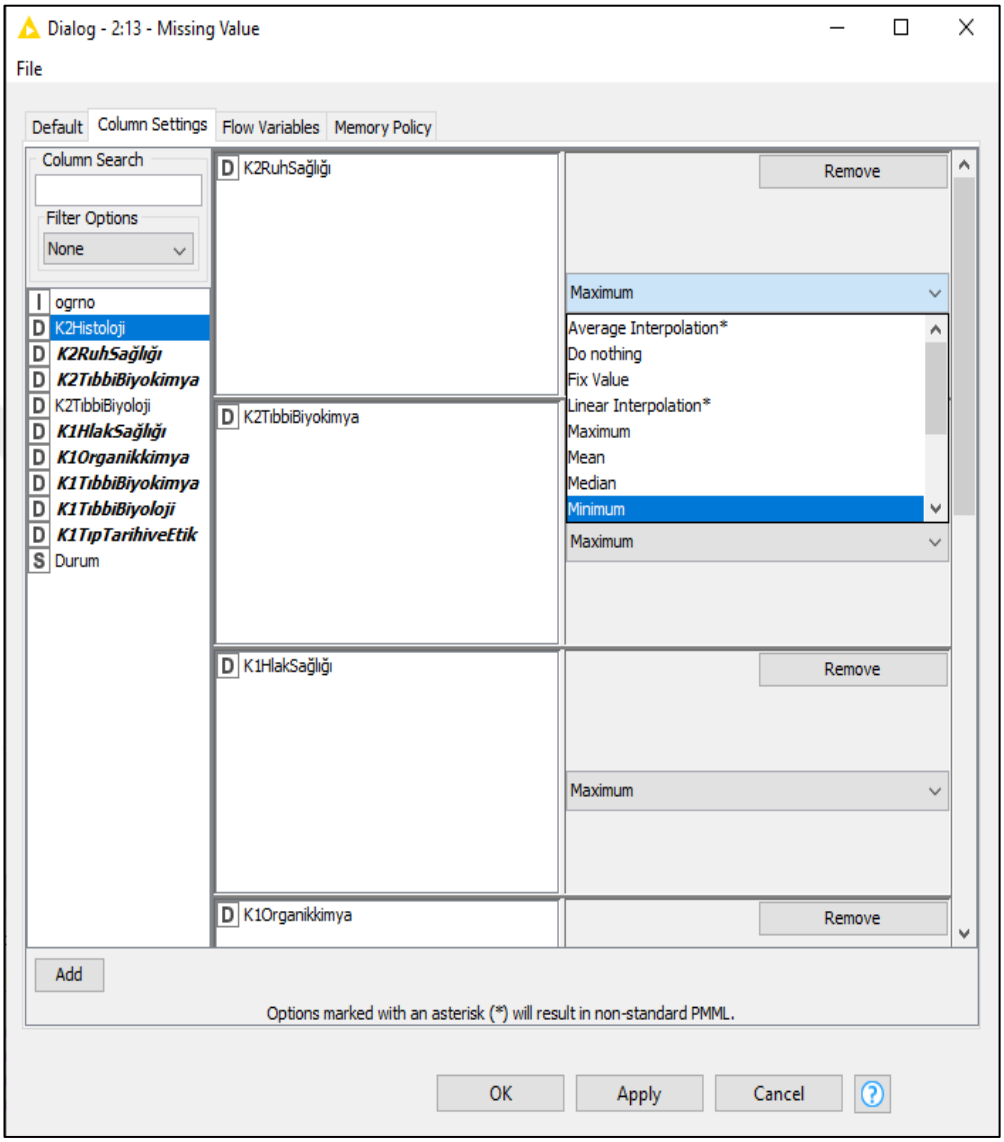
Veri ön işleme aşaması, veri setinin modelleme aşaması için hazırlandığı adımdır. Kaliteli bilgiye ulaşmanın en temel şartı kaliteli veridir. Uygulamanın bu adımında verinin anlaşılması adımında elde edilen analiz sonuçlarından faydalanılmıştır. İlk olarak, eksik veri problemine dönük çalışmalar yapılmıştır. Daha sonra veri normalizasyon işlemi gerçekleştirilmiştir.

3.3.3.1. Eksik veri problemi

Uygulamanın bu adımında, eksik veri problemine dönük çalışmalar gerçekleştirilmiştir. Veri seti içersinde yer alan eksik veriler, verinin anlaşılması adımında Statistics düğümü kullanılarak tespit edilmiştir. Eksik veri probleminin çözümü için farklı yöntemler denenerek oluşturulan model için en başarılı yöntemin belirlenmesine çalışılmıştır.

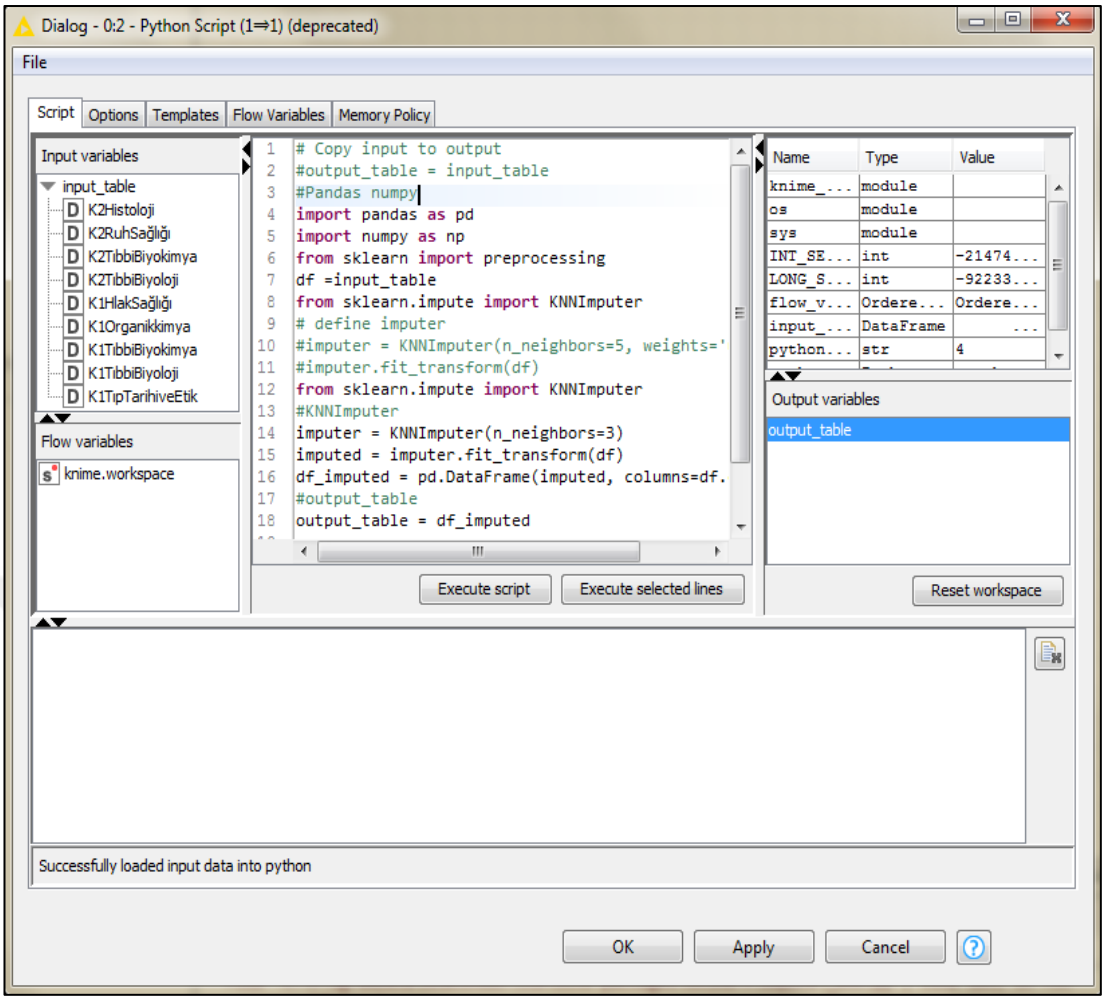
İlk olarak KNIME platformunda yer alan Missing Value düğümü kullanılmıştır. Missing Value düğümünde öncelikle, column search alanından eksik verilerin yer aldığı kolonlar seçilir. Daha sonra eksik veri problemi belirlenen yöntem veya yöntemlere göre çözülür. Eksik veri içeren her kolon için aynı yöntem seçilebileceği gibi, her kolon için farklı yöntem belirlenebilir. Veri setinde tespit edilen eksik veri probleminin çözümü için, missing value düğümünde yer alan mean (ortalama) yöntemi

kullanılmıştır. Mean yöntemi, eksik verinin bulunduğu kolonda yer alan değerlerin aritmetik ortalamasını alarak, eksik verilerin yer aldığı alanları ortalama değeri ile doldurur. Missing Value düğümü yapılandırma ekranı, şekil 3.12'de gösterilmiştir.



Şekil 3.12. Missing Value düğümü yapılandırma ekranı

Eksik veri probleminin çözümü için daha sonra, k-en yakın komşu makine öğrenmesi algoritması kullanılmıştır. Bu yöntemin uygulanması için KNIME veri analiz platformunda yer alan Python Script düğümünden yararlanılmıştır. Python Script düğümü, yerel bir python ortamında bir python betiğinin çalıştırılmasına izin verir. Bu düğüm ile, k-en yakın komşu makine öğrenmesi algoritması kullanılarak, eksik veri probleminin çözümü için python kodları yazılmıştır. Python Script düğümü yapılandırma ekranı şekil 3.13'de gösterilmiştir.

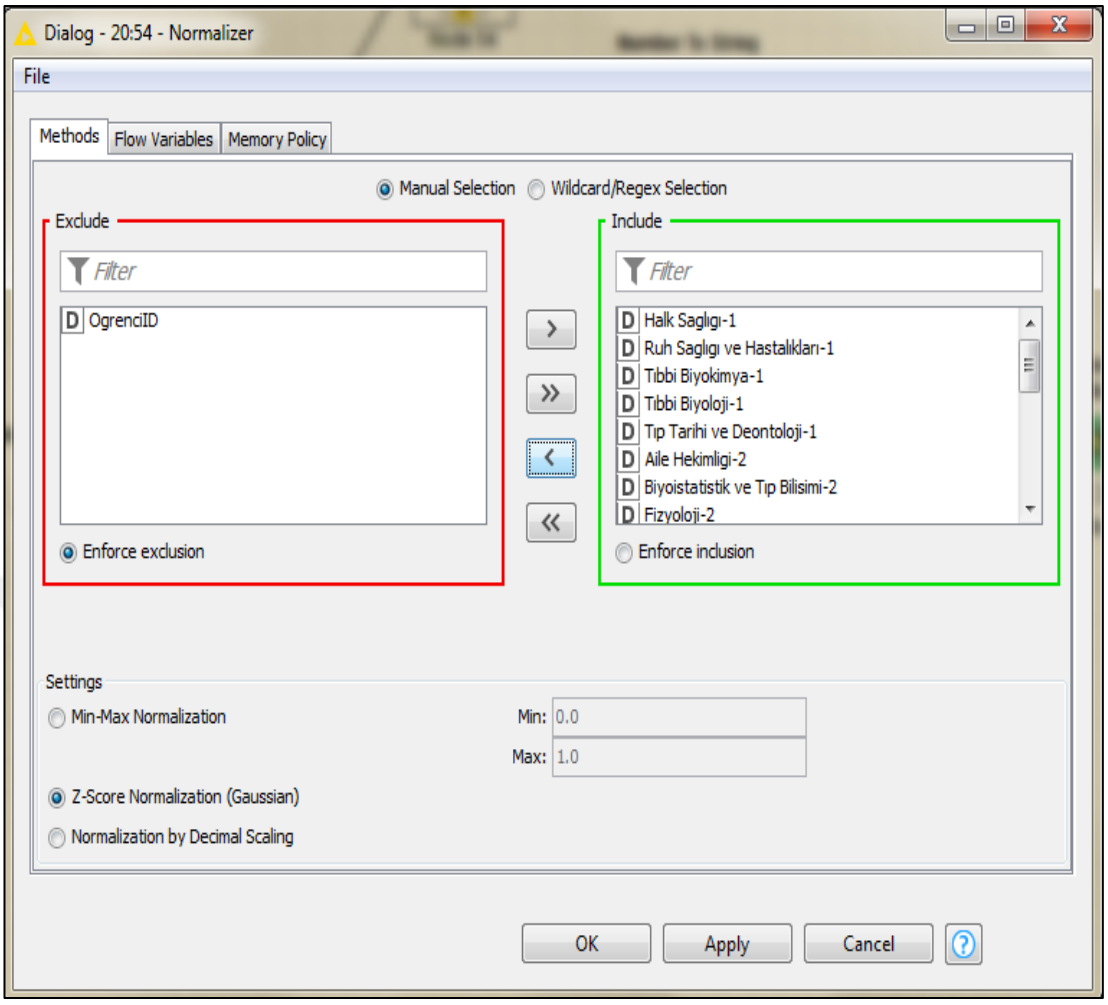


Şekil 3.13. Python Script düğümü yapılandırma ekranı

3.3.3.2. Veri normalizasyonu

Alt kurul derslerine ait soru sayılarının farklılık göstermesi nedeni ile veri seti üzerinde normalizasyon işlemi uygulanarak farklı ölçeklerde ve çok geniş tanım aralıklarına yayılan verilerin, tek bir düzen içinde ifade edilmesi amaçlanmıştır.

Veri normalizasyon işlemi, KNIME veri analiz platformunda yer alan Normalizer düğümü kullanılarak gerçekleştirilmiştir. Çalışmada, veri normalizasyon yöntemi olarak Z-Score normalizasyon yöntemi kullanılmıştır. Normalizer yapılandırma ekranında ilk önce, Exclude alanından normalizasyon işlemi uygulanacak alanlar seçilmiş ve Include alanına aktarılmıştır. Daha sonra, Settings alanından uygulanacak yöntem olarak Z-Score normalizasyon yöntemi seçilerek, belirtilen alanlara ait normalizasyon işlemi gerçekleştirilmiştir. Normalizer düğümü yapılandırma ekranı şekil 3.14'de gösterilmiştir.

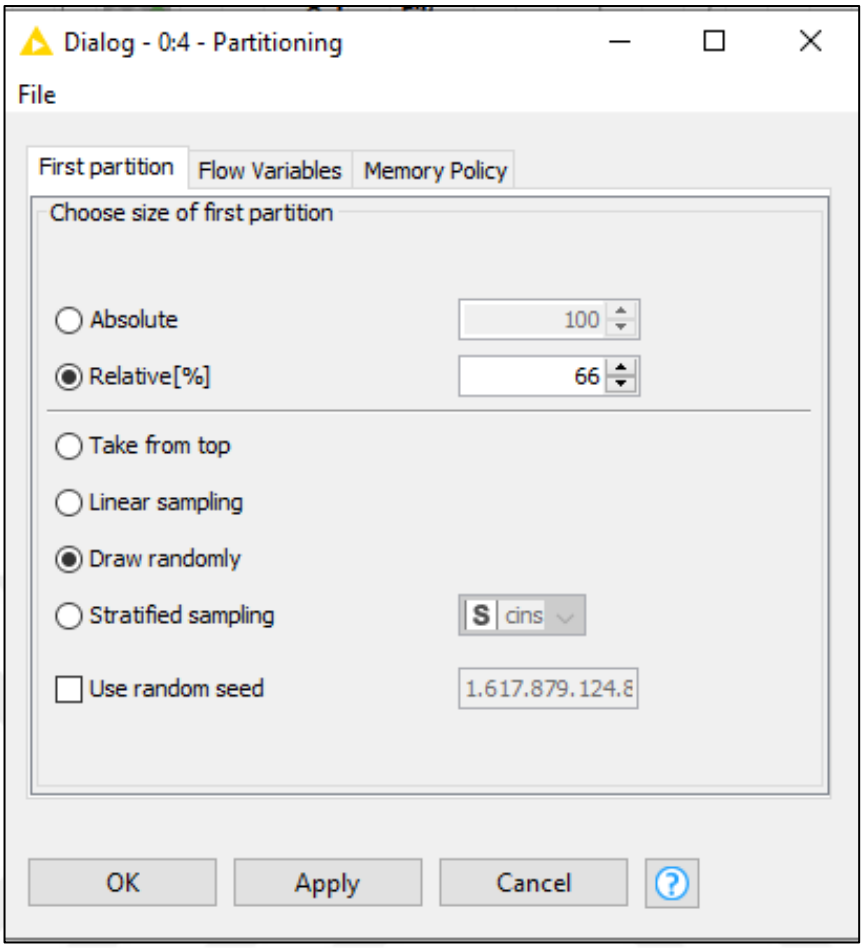


Şekil 3.14. Normalizer düğümü yapılandırma ekranı

3.3.3.3. Model doğrulama yöntemleri

Sınıflandırma yöntemlerinde modelin eğitilmesi ve test edilmesi için veri setinin eğitim ve test veri seti olarak ayrılması gerekmektedir. Çalışma kapsamında veri setinin eğitim ve test veri seti olarak ayrılması işleminde sınama seti yaklaşımı ve k-katlı çapraz doğrulama yöntemleri, kullanılan 5 sınıflandırma yöntemi için uygulanmış ve ilgili yöntemlerin, kullanılan sınıflandırma yöntemlerinin başarı oranlarına etkisi araştırılmıştır.

Çalışmada ilk olarak sınama seti yaklaşımı yöntemi uygulanmıştır. Bu yöntemde veri seti belli oranlarda eğitim ve test veri seti olarak ayrılır. Bu yöntemin uygulanması için Partitioning düğümü kullanılmıştır. Veri setinin 2/3'lük kısmı eğitim, 1/3'lük kısmı test veri seti olarak ayrılmıştır. Partitioning düğümü yapılandırma ekranı şekil 3.15'de gösterilmiş ve yapılandırma ekranına ait açıklamalara aşağıda yer verilmiştir.



Şekil 3.15. Partitioning düğümü yapılandırma ekranı

Absolute: Eğitim veri seti olarak kullanılacak satır sayısının belirtildiği alandır. Girilen değer eğitim veri seti olarak, kalan kısım test veri seti olarak kullanılır.

Relative[%]: Eğitim veri setinin yüzde olarak belirlendiği alandır. Girilen değer eğitim veri seti oranını belirler, kalan kısım test veri seti olarak kullanılır.

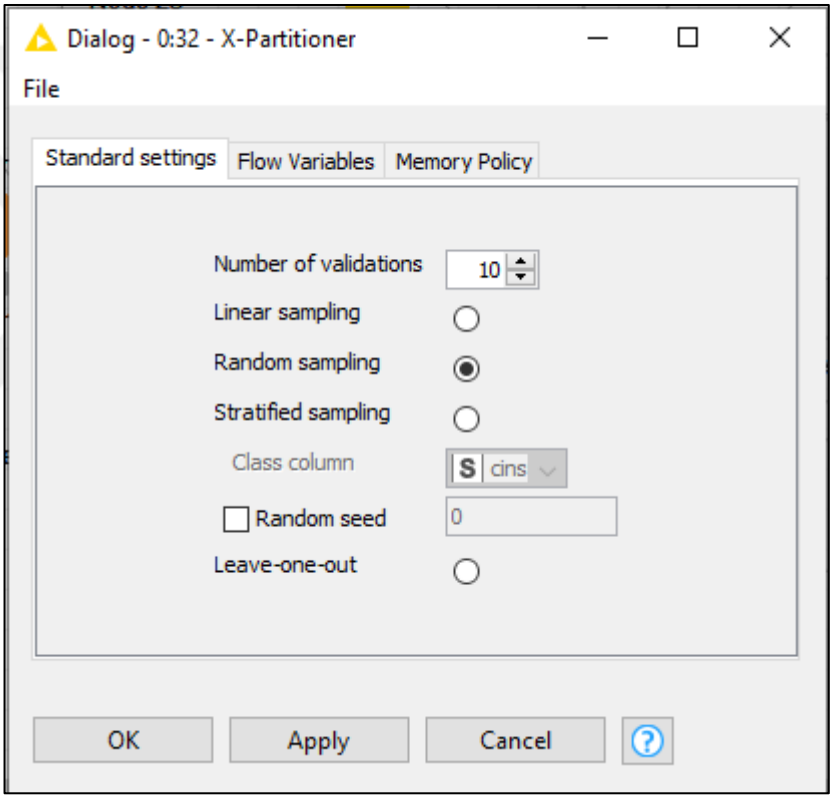
Teke From Top: Veri setini sıralı olarak eğitim ve test verisi olarak ayırır.

Linear Sampling: Tüm satırlar üzerinden doğrusal olarak seçim yapılır.

Draw Randomly: Veri setini rastgele olarak eğitim ve test verisi olarak ayırır.

Daha sonra veri seti, k-katlı çapraz doğrulama yöntemi kullanılarak, eğitim ve test veri seti olarak ayrılmıştır. K-katlı çapraz doğrulama yönteminde veri seti k adet parçaya ayrılır. Oluşan k adet parçadan bir parça test kümesi k-1 adet parça eğitim kümesi olarak kullanılır. Bu işlem tüm veri seti için tekrarlanır. K-katlı çapraz doğrulama

yönteminin uygulanması için, x-partitioner ve x-aggregator düğümleri kullanılmıştır. X-Partitioner düğümü, çapraz doğrulama döngüsündeki ilk düğümdür. Bu iki düğüm arasındaki tüm düğümler, belirlenen yenileme sayısı kadar çalıştırılır. X-Aggregator düğümü, çapraz doğrulama döngüsünün sonunda yer alır. Tahmin sonuçlarını toplar, tahmin edilen sınıf değerleri ile gerçek sınıf değerlerini karşılaştırır ve tüm yenileme adımları için istatistik değerlerini çıkarır. Çalışmada yenileme değeri 10 olarak belirlenmiştir. X-Partitioner düğümü yapılandırma ekranı şekil 3.16'da, X-Aggregator düğümü yapılandırma ekranı şekil 3.17'de gösterilmiş yapılandırma ekranlarına ait açıklamalara aşağıda yer verilmiştir.



Şekil 3.16. X-Partitioner düğümü yapılandırma ekranı

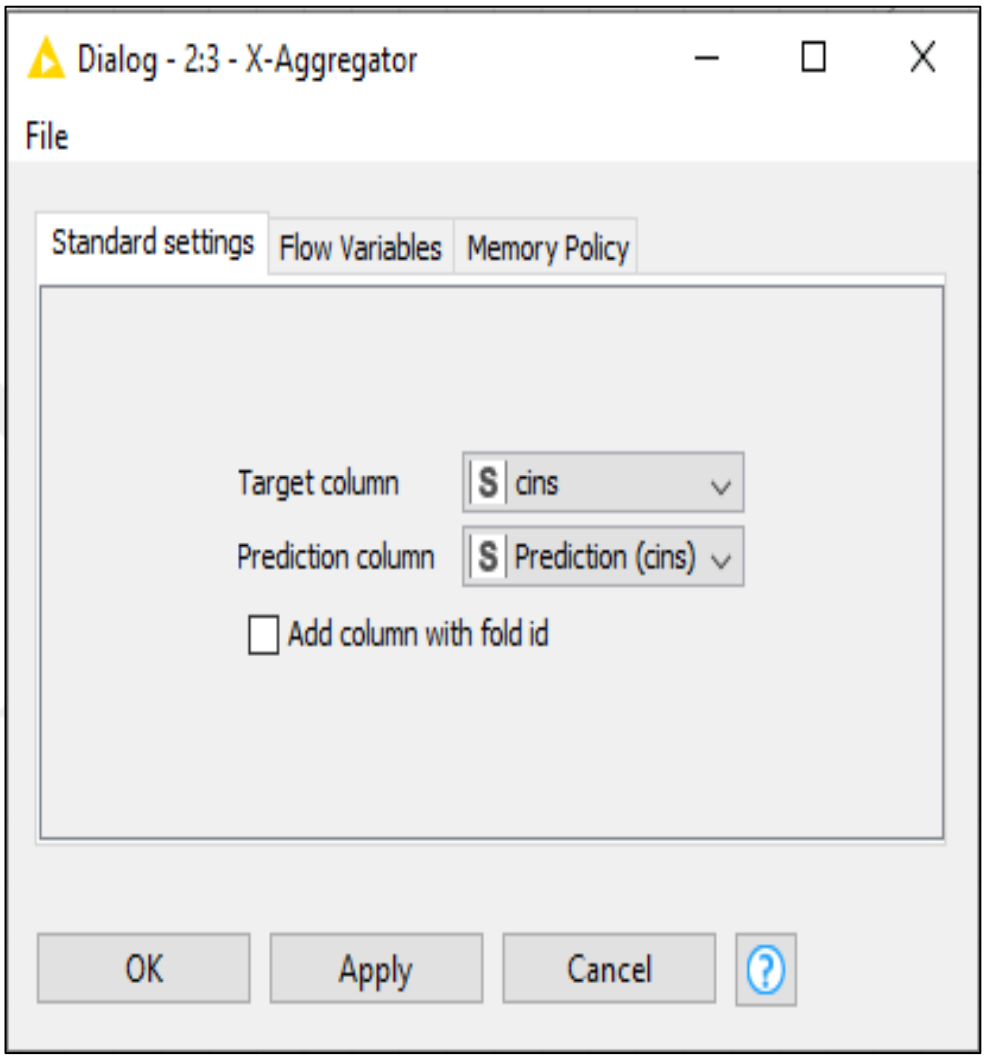
Number of Validation: Gerçekleştirilmesi istenen çapraz doğrulama yineleme sayısının belirlendiği alandır.

Linear Sampling: Giriş verileri sıralı parçalara bölünür.

Random Sampling: Giriş verileri rastgele parçalara bölünür.

Stratified Sampling: Bölümler rastgele örneklenir, ancak Class column alanından seçilen sınıf değeri korunur.

Leave-one-out: Bir örnek dışarda bırakılarak çapraz doğrulama gerçekleştirilir.



Şekil 3.17. X-Aggregator düğümü yapılandırma ekranı

Target Column: Gerçek sınıf etiketini içeren sütun.

Prediction Column: Tahmin etiketini içeren sütun.

3.3.4. Modelleme

Bu aşamada, belirlenen problemin çözümüne yönelik olarak veri ön işleme adımında hazırlanan veri seti üzerinde veri madenciliği yöntemleri uygulanmıştır. Tez çalışmasında, Karar Ağaçları, Yapay Sinir Ağları, Rastgele Orman, K-En Yakın

Komşu ve Naive Bayes olmak üzere 5 farklı veri madenciliği yöntemi uygulanarak belirlenen problemin çözümü için en başarılı yöntem araştırılmıştır.

3.3.4.1. Karar ağaçları yöntemi

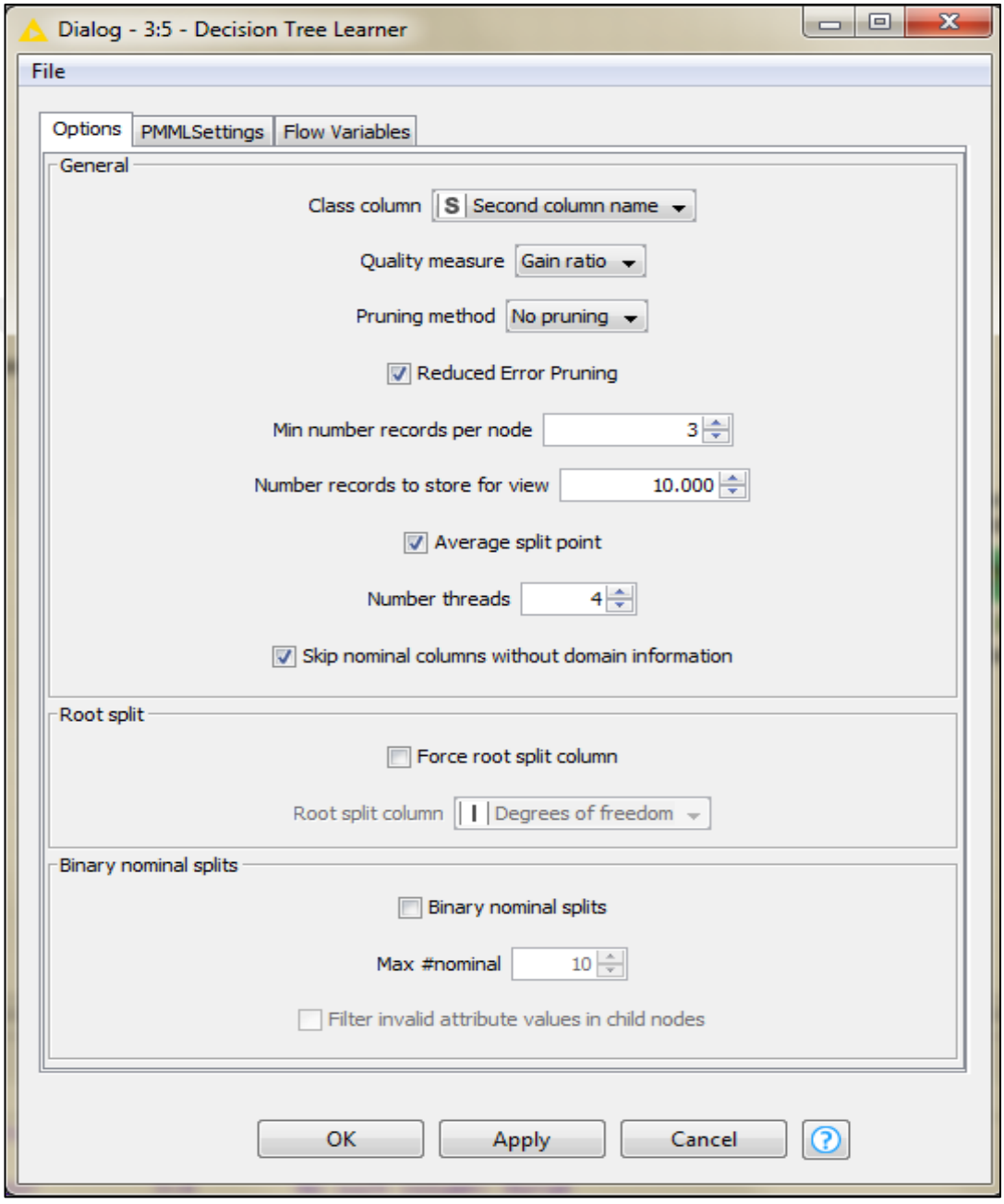
Karar ağaçları, sınıflandırma problemlerinin çözümü için sıklıkla tercih edilen bir veri madenciliği yöntemidir. Düşük maliyetli oluşu, yorumlanması ve anlaşılmasının kolaylığı nedeni ile yaygın olarak kullanılan yöntemlerden bir tanesidir (Chien vd., 2008).

Karar ağaçları yönteminin uygulanması için öncelikle Decision Tree Learner düğümü çalışma ekranına eklenmiştir. Decision Tree Learner düğümü, bir karar ağacı yapısı oluşturur. Bu düğüm, eğitim veri seti kullanılarak modelin eğitilmesini ve kural setinin oluşturulmasını sağlar. Decision Tree Learner düğümüne ait yapılandırmalar şekil 3.18'de yer alan ekran üzerinden gerçekleştirilmiş ve yapılandırma ekranına ait açıklamalara aşağıda yer verilmiştir. Bu düğüm üzerinden elde edilen karar ağacı yapısı, Decision Tree Predictor düğümüne aktarılmıştır. Decision Tree Predictor düğümü, kural seti ve test veri kümesini kullanarak, test veri seti içersindeki gözlemlerin sınıf değerlerini tahmin eder. KNIME veri analiz platformu üzerinde karar ağaçları yöntemi kullanılarak oluşturulan model Şekil 3.19'da gösterilmiştir. Karar ağaçları yönteminin uygulanması sonucu oluşan karışıklık matrisi Tablo 3.3'de, dönem 1 veri setine ait oluşan karar ağacı yapısı şekil 3.20'de gösterilmiştir.

Karar ağaçları yönteminde, dönem 1 öğrencileri için en yüksek doğruluk oranı eksik veri probleminin k-en yakın komşu makine öğrenme algoritması kullanılarak çözüldüğü, model doğrulama yöntemi olarak sınama seti yaklaşımı yönteminin kullanıldığı model ile elde edilmiştir. Bu modelde doğruluk oranı % 90,47 olarak hesaplanmıştır.

Karar ağaçları yönteminde, dönem 2 öğrencileri için en yüksek doğruluk oranı eksik veri probleminin k-en yakın komşu algoritması ve ortalama yöntemi kullanılarak çözüldüğü, model doğrulama yöntemi olarak sınama seti yaklaşımı yönteminin kullanıldığı modeller ile elde edilmiştir. Bu modellerde doğruluk oranı % 93,33 olarak hesaplanmıştır.

Karar ağaçları yönteminde, dönem 3 öğrencileri için en yüksek doğruluk oranı eksik veri probleminin ortalama yöntemi kullanılarak çözüldüğü, model doğrulama yöntemi olarak sınama seti yaklaşımı yönteminin kullanıldığı model ile elde edilmiştir. Bu modelde doğruluk oranı % 91,86 olarak hesaplanmıştır.



Şekil 3.18. Decision Tree Learner düğümü yapılandırma ekranı

Class column: Sınıf etiketi belirlenmek istenen hedef niteliğin seçildiği alandır. Yalnızca nominal değerler içeren öznitelikler seçilebilir.

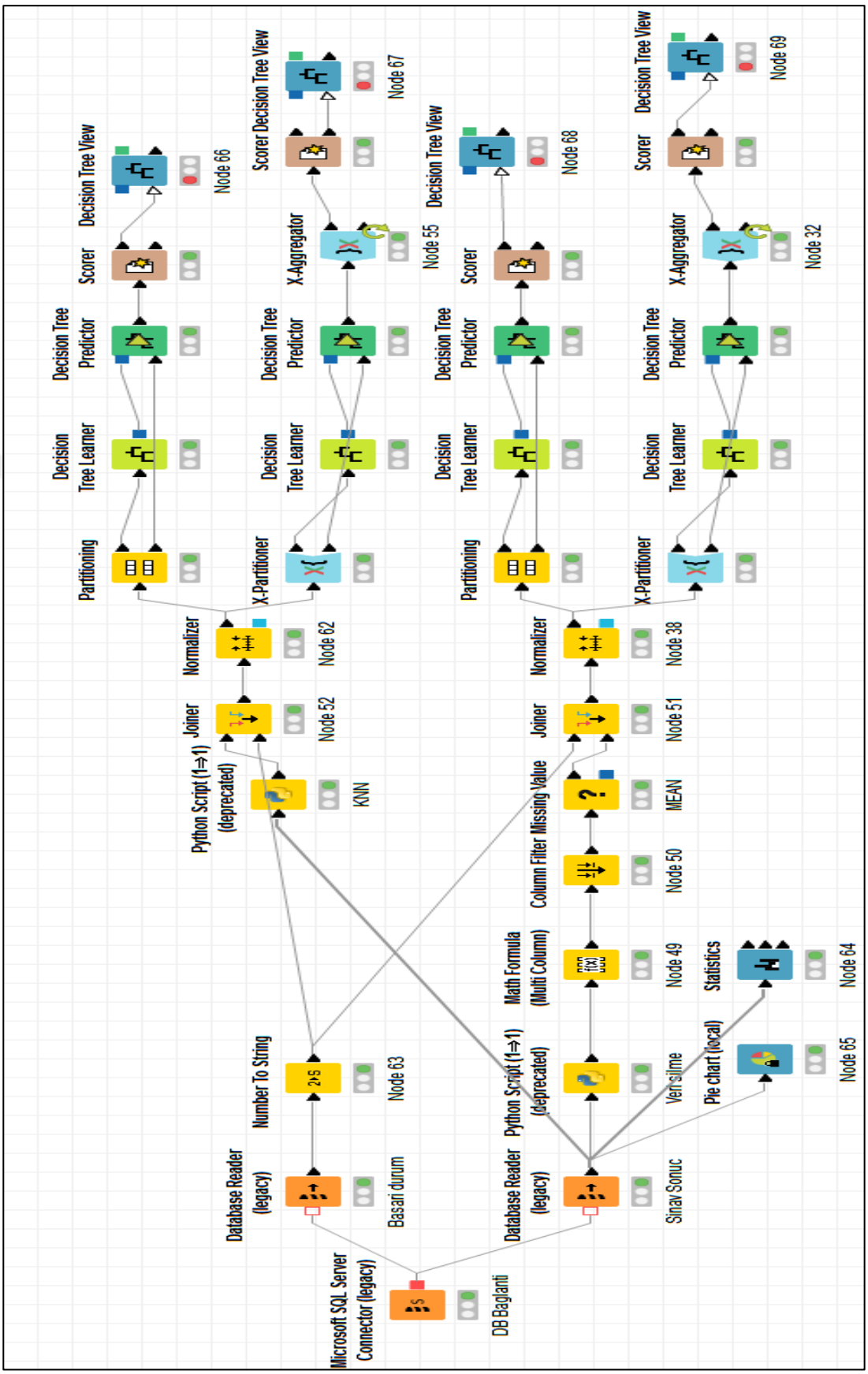
Quality measure: Dallanma yönteminin seçildiği alandır. Gini index ve Gain ratio alanları mevcuttur. Gain ratio yöntemi seçilmiştir.

Pruning method: Karar ağacı budama yönteminin seçildiği alandır. No pruning (budama yok) veya MDL budama yöntemi seçilebilir. No pruning seçilmiştir.

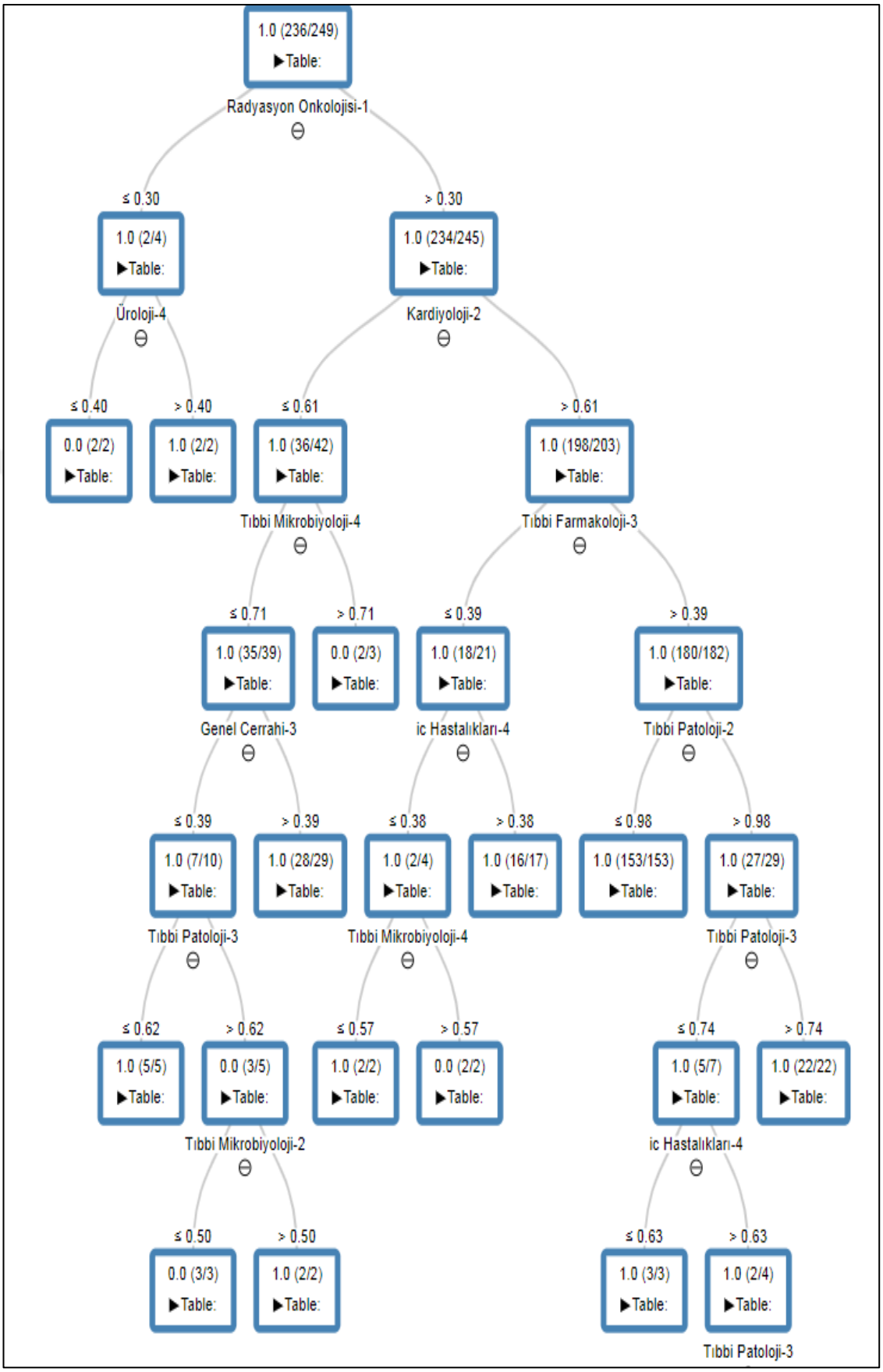
Min number records per node: Her düğümde gereken minimum kayıt sayısının belirlendiği alandır.

Tablo 3.3. Karar Ağaçları yöntemi karışıklık matrisi

Dönem	Eksik Veri Tamamlama Yöntemi	Model Doğrulama Yöntemi	Tahmin Değerleri	Gerçek Değerler		Başarı Oranı (%)
				Başarılı	Başarısız	
Dönem 1	K-En Yakın Komşu	Sınama Seti Yaklaşımı	Başarılı	80	5	90,47
			Başarısız	5	15	
		K-Katlı Çapraz Doğrulama	Başarılı	221	23	84,96
			Başarısız	23	39	
	Ortalama Yöntemi	Sınama Seti Yaklaşımı	Başarılı	72	8	79,04
			Başarısız	14	11	
		K-Katlı Çapraz Doğrulama	Başarılı	222	28	83,66
			Başarısız	22	34	
Dönem 2	K-En Yakın Komşu	Sınama Seti Yaklaşımı	Başarılı	77	4	93,33
			Başarısız	2	7	
		K-Katlı Çapraz Doğrulama	Başarılı	245	12	91,63
			Başarısız	13	29	
	Ortalama Yöntemi	Sınama Seti Yaklaşımı	Başarılı	76	3	93,33
			Başarısız	3	8	
		K-Katlı Çapraz Doğrulama	Başarılı	244	13	90,63
			Başarısız	15	27	
Dönem 3	K-En Yakın Komşu	Sınama Seti Yaklaşımı	Başarılı	86	20	88,77
			Başarısız	9	1	
		K-Katlı Çapraz Doğrulama	Başarılı	255	15	90,21
			Başarısız	13	3	
	Ortalama Yöntemi	Sınama Seti Yaklaşımı	Başarılı	78	4	91,86
			Başarısız	3	1	
		K-Katlı Çapraz Doğrulama	Başarılı	259	11	91,25
			Başarısız	14	2	



Şekil 3.19. Karar Ağaçları yöntemi modeli



Şekil 3.20. Dönem 1 veri setine ait Karar Ağacı yapısı

3.3.4.2. Yapay Sinir Ağları

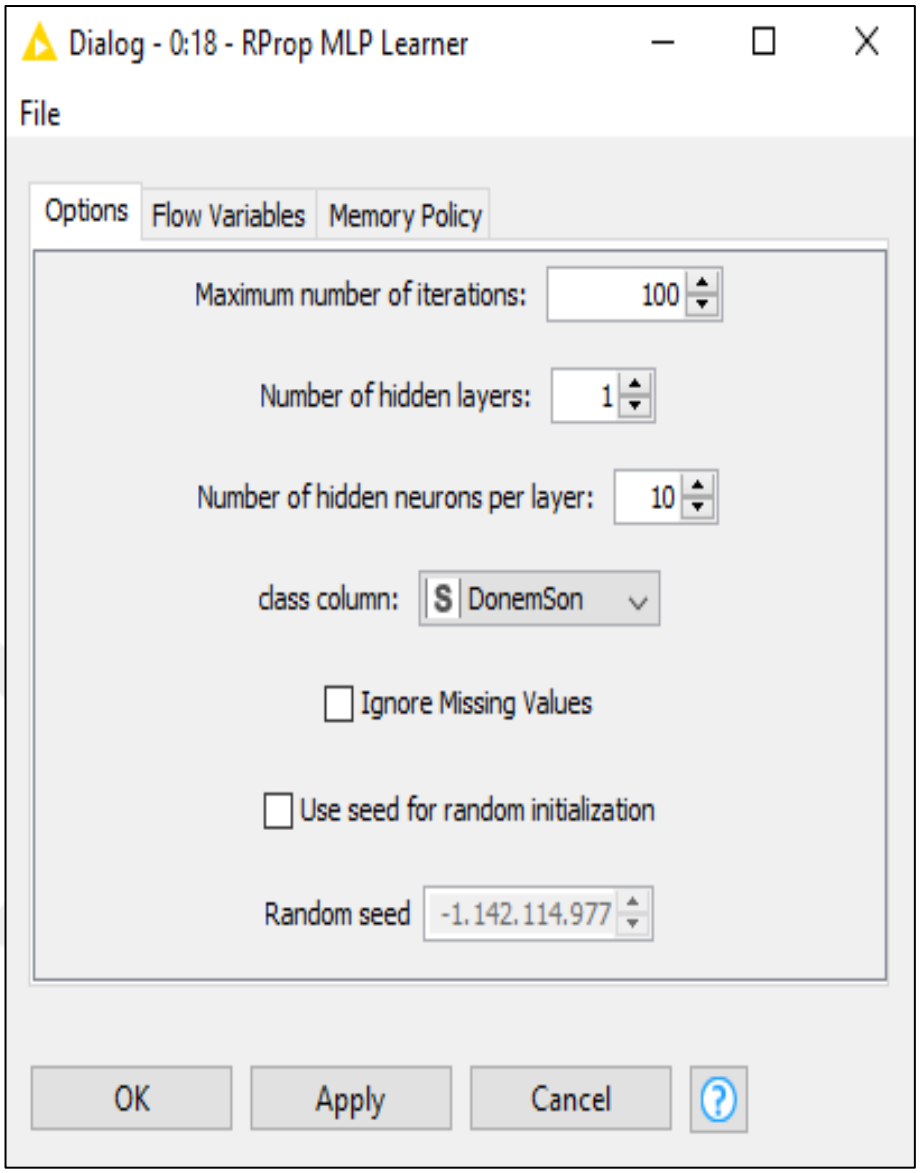
Yapay sinir ağları yöntemi, insan beyninin çalışma prensiplerinden esinlenilerek geliştirlmiş bir ağ modelidir. Nöron hücreleri ve bu hücrelerin bir araya gelerek oluşturdukları yapının, dijital olarak modellenmesidir (Yıldıran vd., 2018). Girdi, ağırlıklar, toplam fonksiyonu, aktivasyon fonksiyonu ve çıktılardan oluşur.

Yapay sinir ağları yönteminin uygulanması için, öncelikle Rprop MLP Learner düğümü çalışma ekranına eklenmiştir. Bu düğüm, eğitim veri seti ile modelin eğitilerek, kural setinin oluşturulmasını sağlar. Rprop MLP Learner düğümüne ait yapılandırma işlemleri, şekil 3.21'de gösterilen ekran üzerinden gerçekleştirilmiştir. Bu düğüm üzerinden elde edilen kural seti, MultiLayer Perceptron Predictor düğümüne aktarılmıştır. MultiLayer Perceptron Predictor düğümü kural seti ve test veri kümesini kullanarak, test veri seti içerisindeki gözlemlerin sınıf değerlerini tahmin eder. Oluşturulan yapay sinir ağları modeli şekil 3.22'de gösterilmiş ve Rprop MLP Learner düğümü yapılandırma ekranına ait açıklamalara aşağıda yer verilmiştir. Yapay sinir ağları yöntemi sonucu oluşan karışıklık matrisi, Tablo 3.4'de gösterilmiştir.

Yapay sinir ağları yönteminde, dönem 1 öğrencileri için en yüksek doğruluk oranı eksik veri probleminin ortalama yöntemi kullanılarak çözüldüğü, model doğrulama yöntemi olarak sınama seti yaklaşımı yönteminin kullanıldığı model ile elde edilmiştir. Bu modelde doğruluk oranı % 91,86 olarak hesaplanmıştır.

Yapay sinir ağları yönteminde, dönem 2 öğrencileri için en yüksek doğruluk oranı eksik veri probleminin k-en yakın komşu makine öğrenme algoritması kullanılarak çözüldüğü, model doğrulama yöntemi olarak k katlı çapraz doğrulama yönteminin kullanıldığı model ile elde edilmiştir. Bu modelde doğruluk oranı % 92,30 olarak hesaplanmıştır.

Yapay sinir ağları yönteminde, dönem 3 öğrencileri için en yüksek doğruluk oranı eksik veri probleminin k-en yakın komşu makine öğrenme algoritması kullanılarak çözüldüğü, model doğrulama yöntemi olarak sınama seti yaklaşımı yönteminin kullanıldığı model ile elde edilmiştir. Bu modelde doğruluk oranı % 96,93 olarak hesaplanmıştır.



Şekil 3.21 Rprop MLP Learner düğümü yapılandırma ekranı

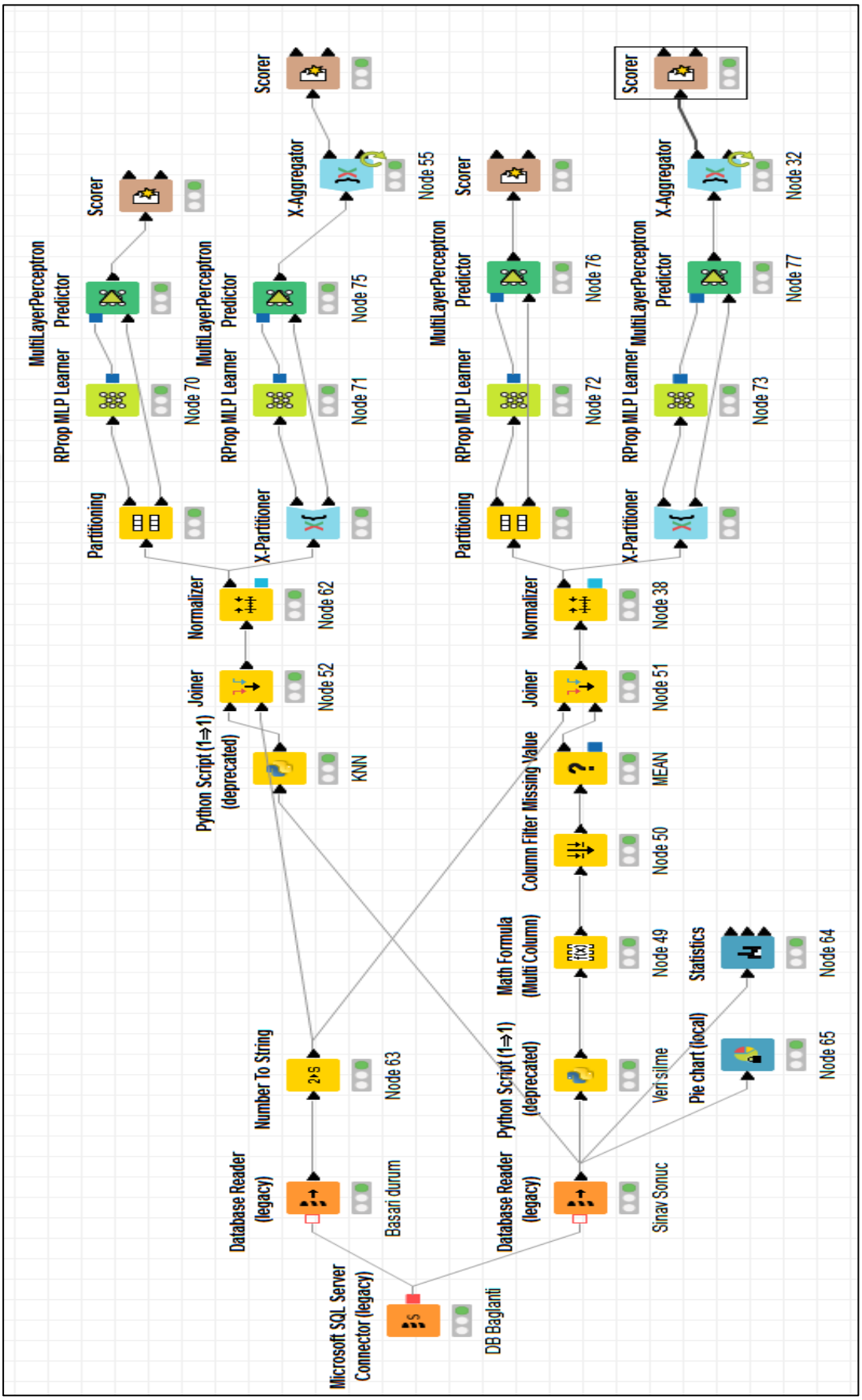
Maximum number of iterations: Öğrenme yenilemelerinin sayısını belirtir.

Number of hidden layers: Yapay sinir ağı yapısındaki gizli katmaların sayısının belirlendiği alandır.

Number of hidden neurons per layer: Her gizli katmanda bulunan nöronların sayısını belirtir.

Class column: Sınıf etiketi belirlenmek istenen hedef niteliğin seçildiği alandır

Ignore missing values: Bu onay kutusu işaretlenirse eksik değerler eğitim için kullanılmayacaktır.



Şekil 3.22. Yapay Sinir Ağları modeli

Tablo 3.4 Yapay Sinir Ağları yöntemi karışıklık matrisi

Dönem	Eksik Veri Tamamlama Yöntemi	Model Doğrulama Yöntemi	Tahmin Değerleri	Gerçek Değerler		Başarı Oranı (%)
				Başarılı	Başarısız	
Dönem 1	K-En Yakın Komşu	Sınama Seti Yaklaşımı	Başarılı	96	2	88,77
			Başarısız	9	1	
		K-Katlı Çapraz Doğrulama	Başarılı	255	15	90,21
			Başarısız	13	3	
	Ortalama Yöntemi	Sınama Seti Yaklaşımı	Başarılı	78	4	91,86
			Başarısız	3	1	
		K-Katlı Çapraz Doğrulama	Başarılı	259	11	91,25
			Başarısız	14	2	
Dönem 2	K-En Yakın Komşu	Sınama Seti Yaklaşımı	Başarılı	87	3	92,15
			Başarısız	5	7	
		K-Katlı Çapraz Doğrulama	Başarılı	250	7	92,30
			Başarısız	16	26	
	Ortalama Yöntemi	Sınama Seti Yaklaşımı	Başarılı	82	5	90,19
			Başarısız	5	10	
		K-Katlı Çapraz Doğrulama	Başarılı	248	16	91,63
			Başarısız	9	26	
Dönem 3	K-En Yakın Komşu	Sınama Seti Yaklaşımı	Başarılı	94	0	96,93
			Başarısız	3	1	
		K-Katlı Çapraz Doğrulama	Başarılı	270	0	94,75
			Başarısız	15	1	
	Ortalama Yöntemi	Sınama Seti Yaklaşımı	Başarılı	87	5	88,77
			Başarısız	6	0	
		K-Katlı Çapraz Doğrulama	Başarılı	267	14	94,05
			Başarısız	3	2	

3.3.4.3. Rastgele orman

Rastgele orman yöntemi, birden fazla karar ağacı yapısı oluşturularak, oluşturulan karar ağaçlarının çoğunluk oyuna göre sınıflandırılmak istenen verinin sınıf değerine karar verilmesi mantığına dayanır.

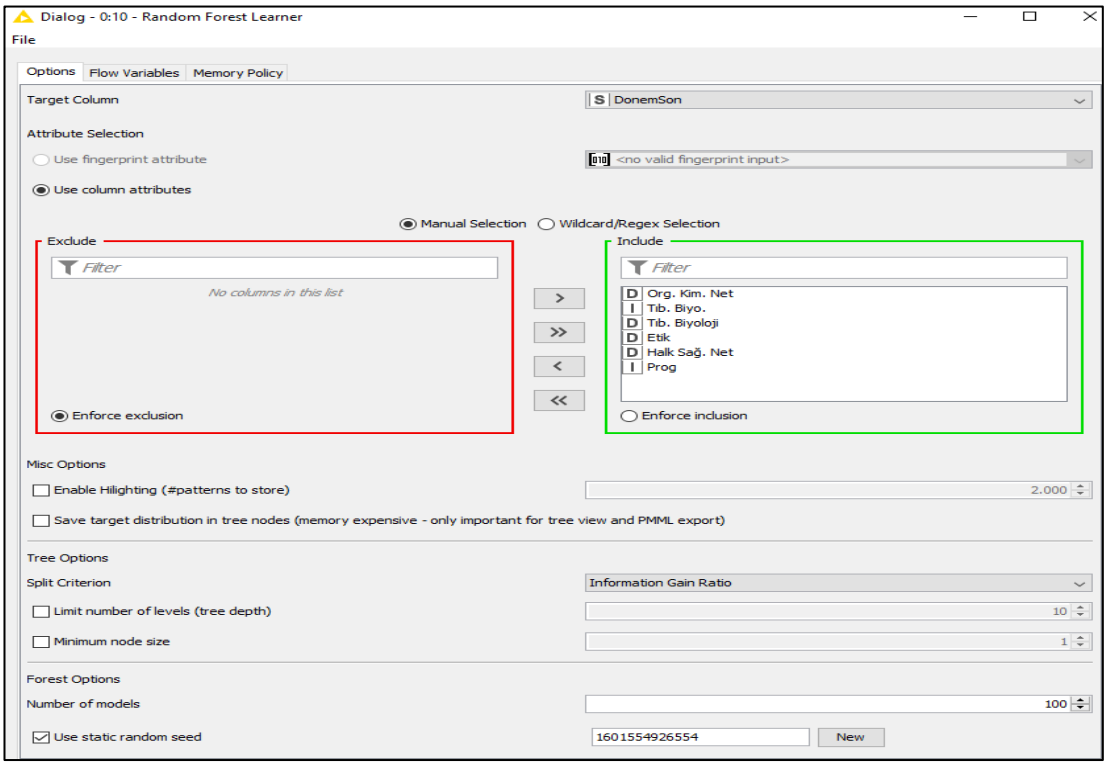
Rastgele orman yönteminin uygulanması için, öncelikle Random Forest Learner düğümü çalışma ekranına eklenmiştir. Bu düğüm, eğitim veri seti ile modelin eğitilerek, kural setinin oluşturulmasını sağlar. Random Forest Learner düğümüne ait yapılandırmalar şekil 3.23'de gösterilen ekran üzerinden gerçekleştirilir. Bu düğüm üzerinden elde edilen kural seti, Random Forest Predictor düğümüne aktarılmıştır. Random Forest Predictor düğümü kural seti ve test veri kümesini kullanarak, test veri seti içerisindeki gözlemlerin sınıf değerlerini tahmin eder. Oluşturulan Rastgele Orman yöntemine ait model Şekil 3.24'de gösterilmiş ve Random Forest Learner

düğümü yapılandırma ekranına ait açıklamalara aşağıda yer verilmiştir. Rastgele orman yöntemi sonucu oluşan karışıklık matrisi, Tablo 3.5'de gösterilmiştir.

Rastgele orman yönteminde, dönem 1 öğrencileri için en yüksek doğruluk oranı eksik veri probleminin k-en yakın komşu makine öğrenme algoritması kullanılarak çözüldüğü, model doğrulama yöntemi olarak k-katlı çapraz doğrulama yönteminin kullanıldığı model ile elde edilmiştir. Bu modelde doğruluk oranı % 91,5 olarak hesaplanmıştır.

Rastgele orman yönteminde, dönem 2 öğrencileri için en yüksek doğruluk oranı eksik veri probleminin ortalama yöntami kullanılarak çözüldüğü, model doğrulama yöntemi olarak sınama seti yaklaşımı yönteminin kullanıldığı model ile elde edilmiştir. Bu modelde doğruluk oranı % 95,09 olarak hesaplanmıştır.

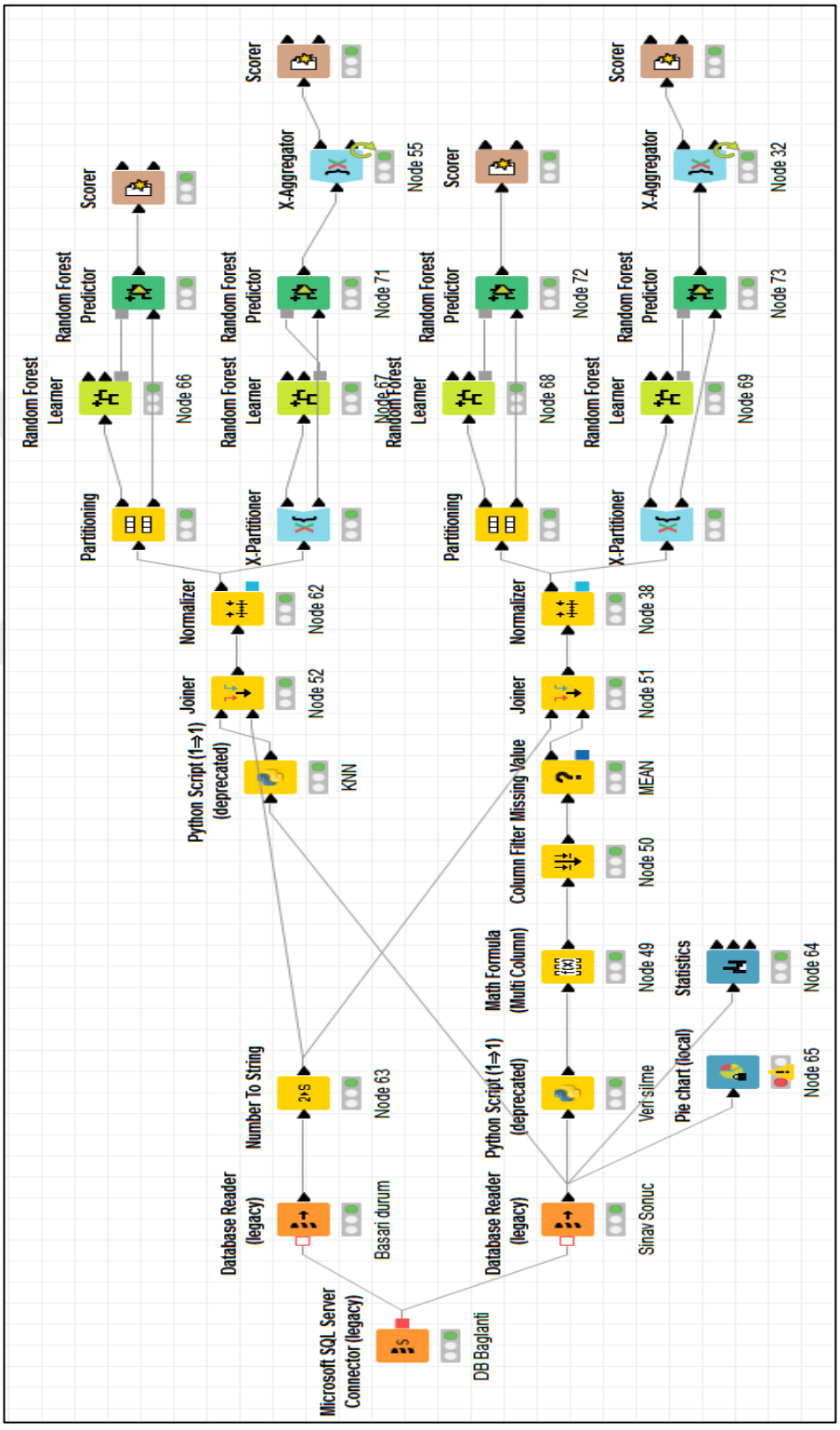
Rastgele orman yönteminde, dönem 3 öğrencileri için en yüksek doğruluk oranı eksik veri probleminin k-en yakın komşu makine öğrenme algoritması kullanılarak çözüldüğü, model doğrulama yöntemi olarak sınama seti yaklaşımı yönteminin kullanıldığı model ile elde edilmiştir. Bu modelde doğruluk oranı % 95,91 olarak hesaplanmıştır.



Şekil 3.23. Random Forest Learner düğümü yapılandırma ekranı

Target Coumn: Sınıf etiketi belirlenmek istenen hedef niteliğin seçildiği alandır.

Split Criterion: Dallanma yöntemi belirlenir. Dallanma yöntemi olarak, Information Gain Ratio yöntemi seçilmiştir.



Şekil 3.24. Rastgele Orman yöntemi modeli

Tablo 3.5. Rastgele Orman yöntemi karışıklık matrisi

Dönem	Eksik Veri Tamamlama Yöntemi	Model Doğrulama Yöntemi	Tahmin Değerleri	Gerçek Değerler		Başarı Oranı (%)
				Başarılı	Başarısız	
Dönem 1	K-En Yakın Komşu	Sınama Seti Yaklaşımı	Başarılı	82	1	91,42
			Başarısız	8	14	
		K-Katlı Çapraz Doğrulama	Başarılı	239	5	91,5
			Başarısız	21	41	
	Ortalama Yöntemi	Sınama Seti Yaklaşımı	Başarılı	79	4	91,42
			Başarısız	5	17	
		K-Katlı Çapraz Doğrulama	Başarılı	238	24	90,19
			Başarısız	6	38	
Dönem 2	K-En Yakın Komşu	Sınama Seti Yaklaşımı	Başarılı	88	2	94,11
			Başarısız	4	8	
		K-Katlı Çapraz Doğrulama	Başarılı	248	9	92,97
			Başarısız	12	30	
	Ortalama Yöntemi	Sınama Seti Yaklaşımı	Başarılı	85	0	95,09
			Başarısız	5	12	
		K-Katlı Çapraz Doğrulama	Başarılı	243	16	89,96
			Başarısız	14	26	
Dönem 3	K-En Yakın Komşu	Sınama Seti Yaklaşımı	Başarılı	93	2	95,91
			Başarısız	2	1	
		K-Katlı Çapraz Doğrulama	Başarılı	269	1	94,75
			Başarısız	14	2	
	Ortalama Yöntemi	Sınama Seti Yaklaşımı	Başarılı	91	1	92,85
			Başarısız	6	0	
		K-Katlı Çapraz Doğrulama	Başarılı	268	15	94,05
			Başarısız	2	1	

3.3.4.4. K-En yakın komşuluk

K-en yakın komşuluk yönteminde, sınıf etiketi belirlenmek istenen gözlem değeri, eğitim veri seti içerisindeki, yeni gözlem değerine en yakın k adet komşunun çoğunluk oyuna göre belirlenir. Uygulamada k değeri, üç olarak belirlenmiştir.

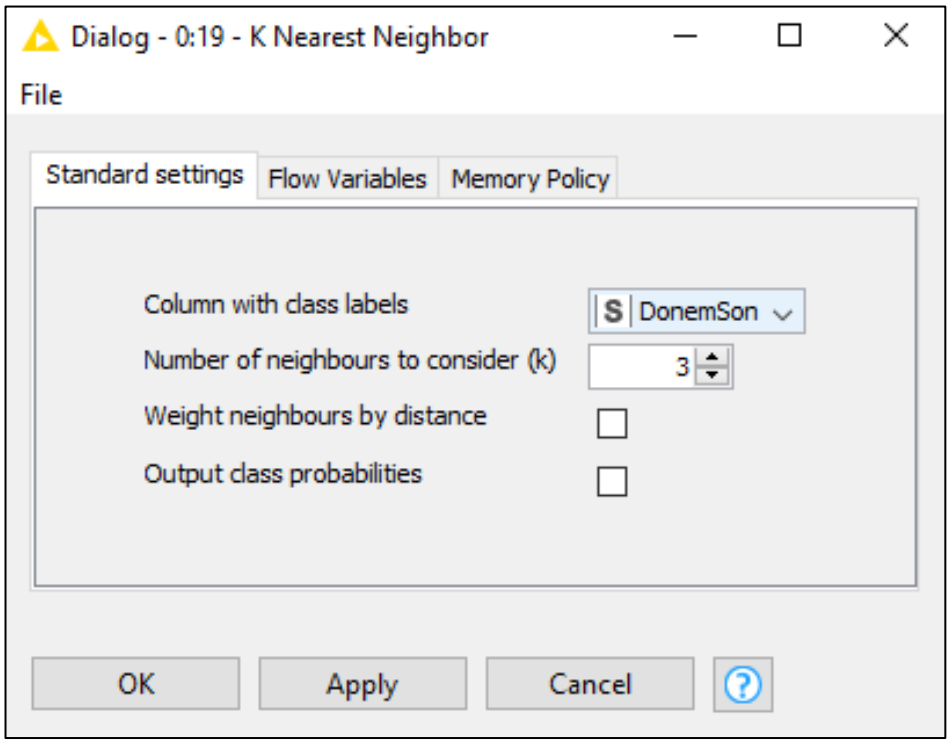
K-en yakın komşuluk yönteminin uygulanması için, öncelikle K Nearest Neighbor düğümü çalışma ekranına eklenmiştir. K-en yakın komşuluk yönteminde, diğer yöntemlerden farklı olarak, eğitim ve test işlemleri aynı düğüm üzerinden gerçekleştirilmektedir. K Nearest Neighbor düğümü eğitim veri seti ile modelin eğitilerek kural setinin oluşturulmasını ve daha sonra test veri seti içerisideki gözlemlerin sınıf etiketlerinin tahmin edilmesini sağlar. K Nearest Neighbor düğümüne ait yapılandırmalar şekil 3.25'de gösterilen ekran üzerinden gerçekleştirilmiş ve yapılandırma ekranına ait açıklamalara aşağıda yer verilmiştir. KNIME üzerinde oluşturulan K-en yakın komşuluk yöntemi modeli Şekil 3.26'da

gösterilmiştir. K-en yakın komşuluk yöntemi sonucu oluşan karışıklık matrisi Tablo 3.6'da gösterilmiştir.

K-en yakın komşuluk yönteminde, dönem 1 öğrencileri için en yüksek doğruluk oranı eksik veri probleminin ortalama yöntemi kullanılarak çözüldüğü, model doğrulama yöntemi olarak sınama seti yaklaşımı yönteminin kullanıldığı model ile elde edilmiştir. Bu modelde doğruluk oranı % 92,38 olarak hesaplanmıştır.

K-en yakın komşuluk yönteminde, dönem 2 öğrencileri için en yüksek doğruluk oranı eksik veri probleminin k-en yakın komşu makine öğrenme algoritması kullanılarak çözüldüğü, model doğrulama yöntemi olarak sınama seti yaklaşımı yönteminin kullanıldığı model ile elde edilmiştir. Bu modelde doğruluk oranı % 92,15 olarak hesaplanmıştır.

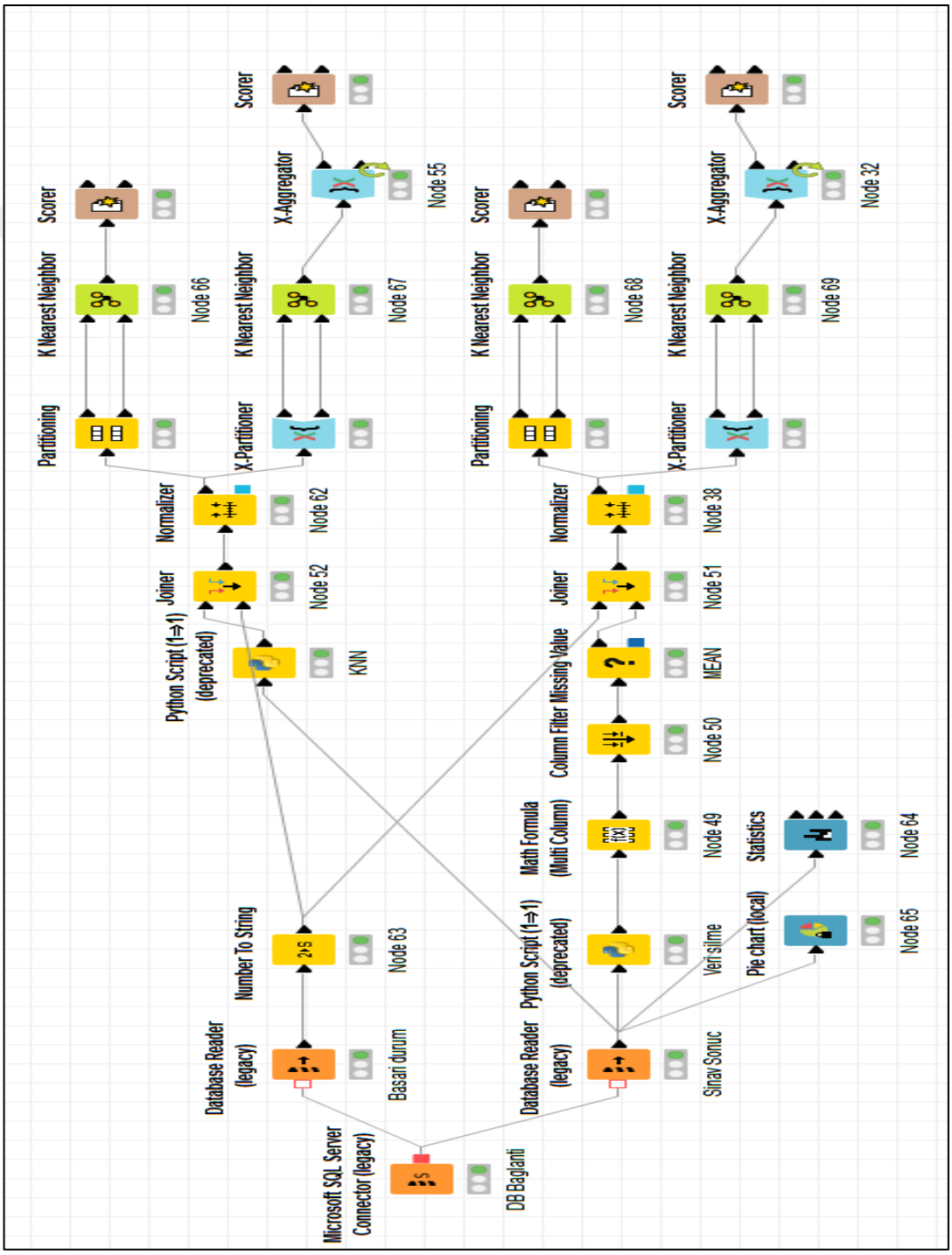
K-en yakın komşuluk yönteminde, dönem 3 öğrencileri için en yüksek doğruluk oranı eksik veri probleminin k-en yakın komşu makine öğrenme algoritması kullanılarak çözüldüğü, model doğrulama yöntemi olarak k-katlı çapraz doğrulama yönteminin kullanıldığı model ile elde edilmiştir. Bu modelde doğruluk oranı % 94,4 olarak hesaplanmıştır.



Şekil 3.25. K Nearest Neighbor düğümü yapılandırma ekranı

Column with class label: Sınıf etiketi belirlenmek istenen, hedef niteliğin seçildiği alandır.

Number of neighbours to consider (k): Sınıflandırılmak istenen örneğin, sınıf etiketinin belirlenmesi için kullanılacak en yakın komşu sayısının belirlendiği alandır. Eşitlik durumu oluşmaması için tek sayı seçilmesi önerilir.



Şekil 3.26. K-en Yakın Komşuluk yöntemi modeli

Tablo 3.6. K-En Yakın Komşuluk yöntemi karışıklık matrisi

Dönem	Eksik Veri Tamamlama Yöntemi	Model Doğrulama Yöntemi	Tahmin Değerleri	Gerçek Değerler		Başarı Oranı (%)
				Başarılı	Başarısız	
Dönem 1	K-En Yakın Komşu	Sınama Seti Yaklaşımı	Başarılı	82	1	88,8
			Başarısız	9	13	
		K-Katlı Çapraz Doğrulama	Başarılı	242	26	90,85
			Başarısız	2	36	
	Ortalama Yöntemi	Sınama Seti Yaklaşımı	Başarılı	86	1	92,38
			Başarısız	7	11	
		K-Katlı Çapraz Doğrulama	Başarılı	241	3	88,88
			Başarısız	31	31	
Dönem 2	K-En Yakın Komşu	Sınama Seti Yaklaşımı	Başarılı	86	2	92,15
			Başarısız	6	8	
		K-Katlı Çapraz Doğrulama	Başarılı	248	9	90,3
			Başarısız	20	22	
	Ortalama Yöntemi	Sınama Seti Yaklaşımı	Başarılı	83	6	88,23
			Başarısız	6	7	
		K-Katlı Çapraz Doğrulama	Başarılı	243	14	87,96
			Başarısız	22	20	
Dönem 3	K-En Yakın Komşu	Sınama Seti Yaklaşımı	Başarılı	91	1	92,85
			Başarısız	6	0	
		K-Katlı Çapraz Doğrulama	Başarılı	270	0	94,4
			Başarısız	6	0	
	Ortalama Yöntemi	Sınama Seti Yaklaşımı	Başarılı	89	1	90,81
			Başarısız	8	0	
		K-Katlı Çapraz Doğrulama	Başarılı	268	2	93,7
			Başarısız	16	0	

3.3.4.5. Naive bayes

Naive Bayes yöntemi, bayes teoremini kullanarak istatistiksel bir çıkarım yapar. Gözlemlerin sınıflara üye olma olasılığını belirlemeyi amaçlar. Her gözlemin sınıflara üye olma olasılığı hesaplanır ve gözlemin üye olma olasılığı en yüksek olan sınıf, gözlemin sınıf değeri olarak belirlenir.

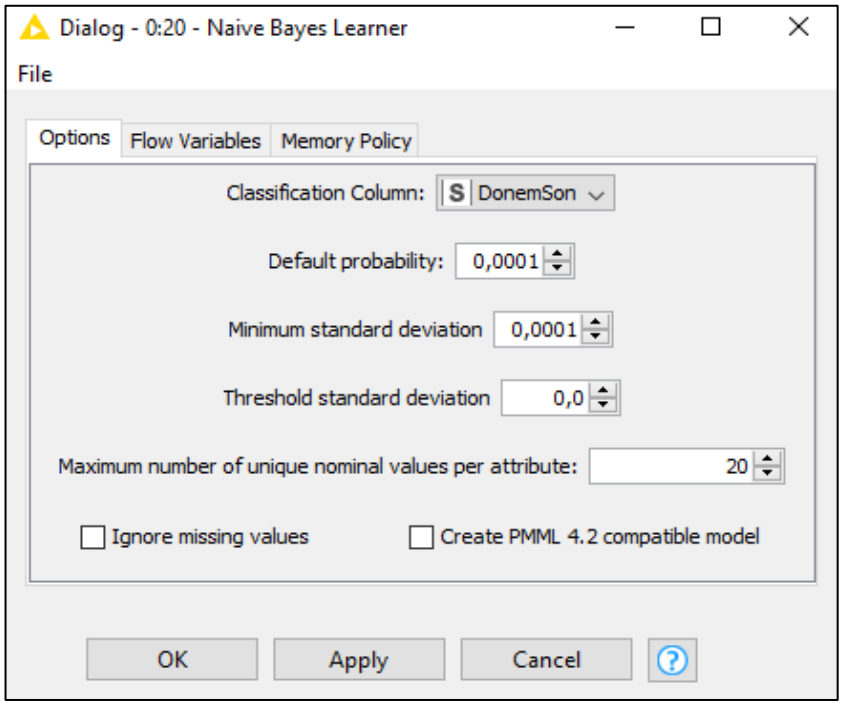
Naive Bayes yönteminin uygulanması için, öncelikle Naive Bayes Learner düğümü çalışma ekranına eklenmiştir. Bu düğüm, eğitim veri seti ile modelin eğitilerek, kural setinin oluşturulmasını sağlar. Naive Bayes Learner düğümüne ait yapılandırmalar şekil 3.27'de gösterilen ekran üzerinden gerçekleştirilmiş ve yapılandırma ekranına ait

açıklamalara aşağıda yer verilmiştir. Bu düğüm üzerinden elde edilen kural seti, Naive Bayes Predictor düğümüne aktarılır. Naive Bayes Predictor düğümü kural seti ve test veri kümesini kullanarak, test veri seti içersindeki gözlemlerin sınıf değrlerini tahmin eder. Oluşturulan Naive Bayes modeli Şekil 3.28'de gösterilmiştir. Naive Bayes yöntemi sonucu oluşan karışıklık matrisi tablosu Tablo 3.7'de gösterilmiştir.

Naive bayes yönteminde, dönem 1 öğrencileri için en yüksek doğruluk oranı eksik veri probleminin k-en yakın komşu makine öğrenme algoritması kullanılarak çözüldüğü, model doğrulama yöntemi olarak sınama seti yaklaşımı yönteminin kullanıldığı model ile elde edilmiştir. Bu modelde doğruluk oranı % 91,42 olarak hesaplanmıştır.

Naive bayes yönteminde, dönem 2 öğrencileri için en yüksek doğruluk oranı eksik veri probleminin ortalama yöntemi kullanılarak çözüldüğü, model doğrulama yöntemi olarak sınama seti yaklaşımı yönteminin kullanıldığı model ile elde edilmiştir. Bu modelde doğruluk oranı % 91,17 olarak hesaplanmıştır.

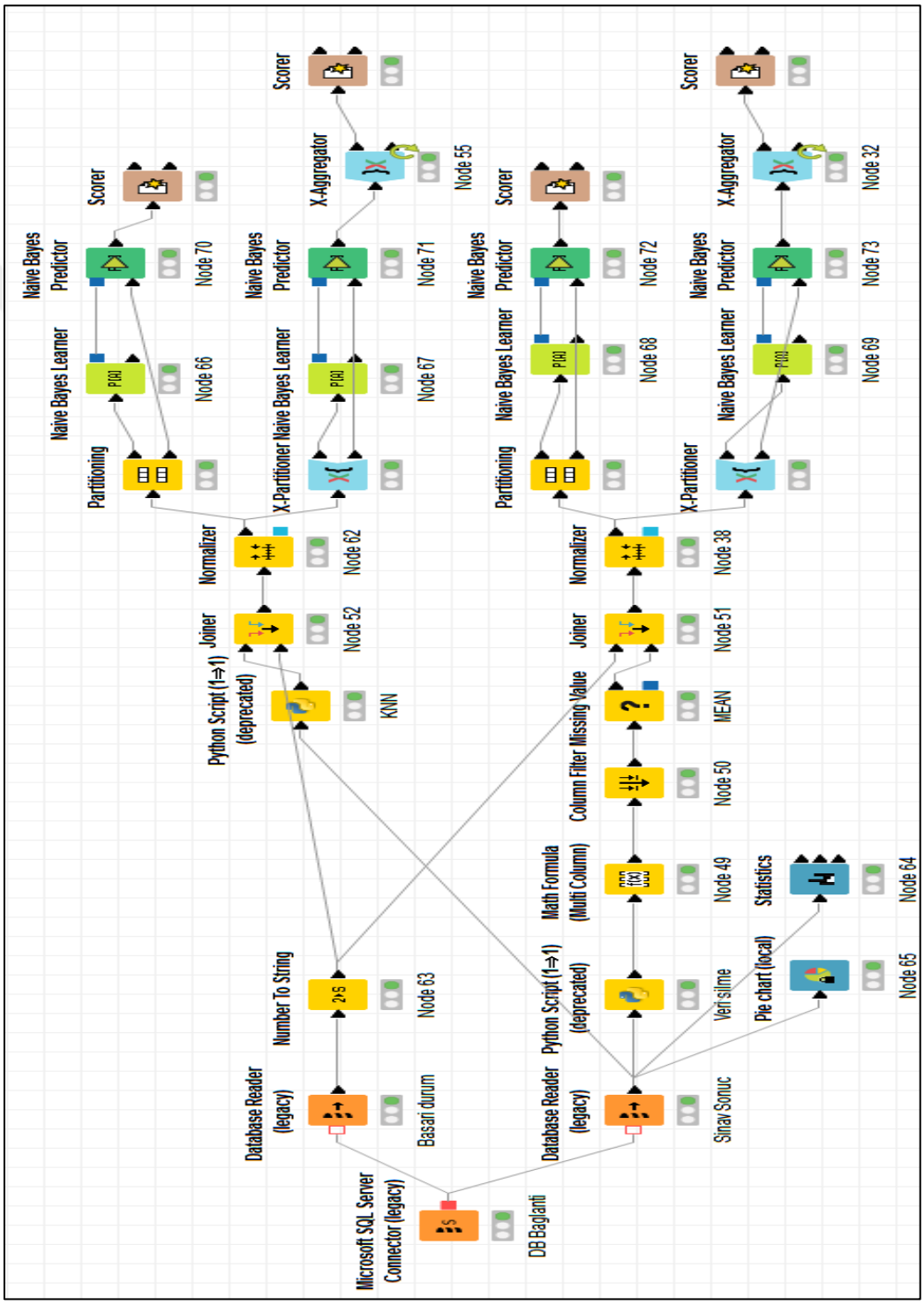
Naive bayes yönteminde, dönem 3 öğrencileri için en yüksek doğruluk oranı eksik veri probleminin k-en yakın komşu makine öğrenme algoritması kullanılarak çözüldüğü, model doğrulama yöntemi olarak sınama seti yaklaşımı yönteminin kullanıldığı model ile elde edilmiştir. Bu modelde doğruluk oranı % 91,86 olarak hesaplanmıştır.



Şekil 3.27. Naive Bayes Learner düğümü yapılandırma ekranı

Classification Column: Sınıf etiketi olarak belirlenmek istenen, hedef niteliğin seçildiği alandır.

Default probability: Olasılık değerinin sıfır olması durumunda kullanılacak değer belirlenir.



Şekil 3.28. Naive Bayes yöntemi modeli

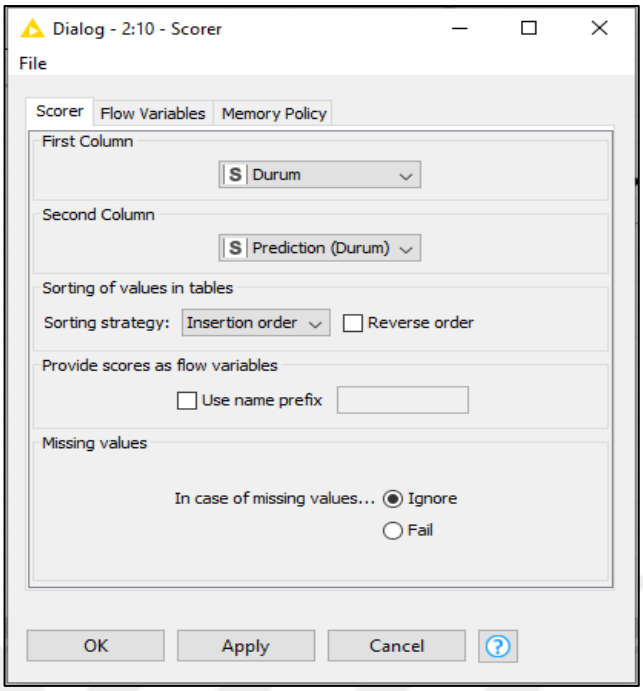
Tablo 3.7. Naive Bayes Yöntemi Karışıklık Matrisi

Dönem	Eksik Veri Tamamlama Yöntemi	Model Doğrulama Yöntemi	Tahmin Değerleri	Gerçek Değerler		Başarı Oranı (%)
				Başarılı	Başarısız	
Dönem 1	K-En Yakın Komşu	Sınama Seti Yaklaşımı	Başarılı	81	7	91,42
			Başarısız	2	15	
		K-Katlı Çapraz Doğrulama	Başarılı	225	19	87,58
			Başarısız	19	43	
	Ortalama Yöntemi	Sınama Seti Yaklaşımı	Başarılı	85	4	90,47
			Başarısız	6	10	
		K-Katlı Çapraz Doğrulama	Başarılı	230	14	89,54
			Başarısız	18	44	
Dönem 2	K-En Yakın Komşu	Sınama Seti Yaklaşımı	Başarılı	72	11	88,23
			Başarısız	1	18	
		K-Katlı Çapraz Doğrulama	Başarılı	230	27	89,96
			Başarısız	3	39	
	Ortalama Yöntemi	Sınama Seti Yaklaşımı	Başarılı	81	8	91,17
			Başarısız	1	12	
		K-Katlı Çapraz Doğrulama	Başarılı	234	23	89,63
			Başarısız	8	34	
Dönem 3	K-En Yakın Komşu	Sınama Seti Yaklaşımı	Başarılı	76	6	91,86
			Başarısız	1	3	
		K-Katlı Çapraz Doğrulama	Başarılı	237	33	86,36
			Başarısız	6	10	
	Ortalama Yöntemi	Sınama Seti Yaklaşımı	Başarılı	85	8	89,79
			Başarısız	2	3	
		K-Katlı Çapraz Doğrulama	Başarılı	237	33	87,8
			Başarısız	6	10	

3.3.5. Değerlendirme

Tez çalışmasının bu aşamasında elde edilen sonuçlar ışığında veri madenciliği sürecinin genel bir değerlendirilmesi yapılmıştır.

Oluşturulan modellerin başarı oranlarının değerlendirilmesinde scorer düğümü kullanılmıştır. Scorer düğümü gerçek değerler ile tahmin değerlerini karşılaştırarak, oluşturulan modele ait doğruluk, hata oranı, kesinlik, hassasiyet ve f-ölçütü değerlerini elde etmemize olanak sağlar. Elde edilen bu değerlerden oluşturulan veri madenciliği modellerinin değerlendirilmesi aşamasında faydalanılır. Scorer düğümü yapılandırma ekranı Şekil 3.29'da gösterilmiştir. Scorer düğümü yapılandırma ekranı kullanılarak Scorer düğümüne ait ayarlar yapılmaktadır. Scorer düğümü üzerinden elde edilen doğruluk, hata oranı, kesinlik, hassasiyet ve f-ölçütü değerleri Tablo 3.8'de yer almaktadır.



Şekil 3.29. Scorer düğümü yapılandırma ekranı

Eksik veri probleminin çözümüne dönük olarak KNIME veri analiz platformunun sunduğu çözümlerin yanı sıra, KNIME platformunda yer alamayan ancak literatürde sıklıkla tercih edilen k-en yakın komşu algoritması kullanılmıştır. Eksik veri probleminin çözümüne dönük olarak kullanılan yöntemlerin oluşturulan veri madenciliği modellerinin başarı oranına etkileri incelendiğinde, dönem 1 öğrencilerine ait veri seti kullanılarak oluşturulan modeller için karar ağaçları, rastgele orman ve naive bayes yöntemleri için k-en yakın komşu makine öğrenme algoritmasının eksik veri probleminin çözümünde daha başarılı sonuçlar verdiği görülmüştür.Yapay sinir ağları ve k-en yakın komşuluk yöntemleri kullanılarak oluşturulan modeler için, eksik veri probleminin ortalama yöntemi kullanılarak çözüldüğü modeler daha başarılı sonuçlar vermiştir.

Dönem 2 öğrencilerine ait veri seti kullanılarak oluşturulan modeller için, karar ağaçları, k-en yakın komşuluk ve yapay sinir ağları yöntemleri için eksik veri probleminin çözümünde k-en yakın komşu makine öğrenme algoritmasının daha başarılı sonuçlar verdiği görülmüştür. Rastgele orman ve naive bayes yöntemleri kullanılarak oluşturulan modeler için eksik veri probleminin ortalama yöntemi kullanılarak çözüldüğü modeler daha başarılı sonuçlar vermiştir.

Dönem 3 öğrencilerine ait veri seti kullanılarak oluşturulan modeller için, naive bayes, k-en yakın komşuluk, rastgele orman ve yapay sinir ağları yöntemleri için eksik veri probleminin çözümünde k-en yakın komşu makine öğrenme algoritmalarının daha başarılı sonuçlar verdiği görülmüştür. Karar ağaçları yöntemi kullanılarak oluşturulan model için eksik veri probleminin ortalama yöntemi kullanılarak çözüldüğü model daha başarılı sonuç vermiştir.

Veri setinin eğitim ve test veri seti olarak ayrılmasında sınama seti yaklaşımı ve k katlı çapraz doğrulama yöntemleri kullanılmıştır. Bu iki yöntemin oluşturulan modellerin başarı oranlarına etkisi değerlendirildiğinde, dönem 1 öğrencilerine ait veri seti kullanılarak oluşturulan modeller için, karar ağaçları, k-en yakın komşuluk, naive bayes ve yapay sinir ağları yöntemleri için model doğrulama yöntemi olarak sınama seti yaklaşımının daha başarılı sonuçlar verdiği görülmüştür. Rastgele orman yöntemi kullanılarak oluşturulan model için doğrulama yöntemi olarak k-katlı çapraz doğrulama yönteminin kullanıldığı model daha başarılı sonuçlar vermiştir.

Dönem 2 öğrencilerine ait veri seti kullanılarak oluşturulan modeller için, karar ağaçları, k-en yakın komşuluk, naive bayes ve rastgele orman yöntemleri için sınama seti yaklaşımı model doğrulama yönteminin daha başarılı sonuçlar verdiği görülmüştür. Yapay sinir ağları yöntemi kullanılarak oluşturulan model için doğrulama yöntemi olarak k-katlı çapraz doğrulama yönteminin kullanıldığı model daha başarılı sonuçlar vermiştir.

Dönem 3 öğrencilerine ait veri seti kullanılarak oluşturulan modeller için, karar ağaçları, rastgele orman, naive bayes ve yapay sinir ağları yöntemleri için sınama seti yaklaşımı model doğrulama yönteminin daha başarılı sonuçlar verdiği görülmüştür. K-en yakın komşuluk yöntemi kullanılarak oluşturulan model için doğrulama yöntemi olarak k-katlı çapraz doğrulama yönteminin kullanıldığı model daha başarılı sonuçlar vermiştir.

Oluşturulan veri madenciliği moldelleri doğruluk oranları değerlendirildiğinde, en yüksek doğruluk oranları dönem 1 öğrencileri için % 92,38, dönem 2 öğrencileri için % 95,09 ve dönem 3 öğrencileri için %96,93 olarak hesaplanmıştır. Yüksek doğruluk oranları göz önüne alındığında öğrencilere ait akademik başarı durumlarının, kurul derslerine ait sınav sonuç verileri kullanılarak erken dönemlerde tahimin

edilebilceği, ayrıca başarılı ve başarısız öğrenci sayılarına dönük çıkarımlarda bulunulabilceği saptanmıştır.

Tablo 3.8. Yöntemlerin değerledirme istatistikleri

Dönem	Sınıflandırma Yöntemi	Eksik Veri Tamamlama Yöntemi	Model Doğrulama Yöntemi	Doğruluk (%)	Hata Oranı (%)	Duyarlılık (%)	Kesinlik (%)	F-Ölçütü (%)
Dönem 1	Karar Ağaçları	Ortalama	Sınama Seti Yaklaşımı	79,04	20,95	57,19	44	50
			K-Katlı Çapraz Doğ.	83,66	16,34	54,8	60,7	57,6
		K-En Yakın Komşu	Sınama Seti Yaklaşımı	**90,4**	9,5	75	75	75
			K-Katlı Çapraz Doğ.	84,96	15,03	62,9	62,9	62,9
	Yapay Sinir Ağları	Ortalama	Sınama Seti Yaklaşımı	**91,86**	8,14	95,1	96,3	95,7
			K-Katlı Çapraz Doğ.	91,25	8,74	95,9	94,9	95,4
		K-En Yakın Komşu	Sınama Seti Yaklaşımı	88,77	11,22	97,7	90,5	94
			K-Katlı Çapraz Doğ.	90,21	9,79	94,4	95,1	94,8
	Random Forest	Ortalama	Sınama Seti Yaklaşımı	**91,42**	8,57	95,2	94	94,6
			K-Katlı Çapraz Doğ.	90,19	9,80	90,8	97,5	94,1
		K-En Yakın Komşu	Sınama Seti Yaklaşımı	**91,42**	8,57	98,8	91,1	94,8
			K-Katlı Çapraz Doğ.	91,5	8,49	98	91,9	94,8
	K En Yakın Komşu	Ortalama	Sınama Seti Yaklaşımı	**92,38**	7,61	98,9	92,5	95,6
			K-Katlı Çapraz Doğ.	88,88	11,11	98,8	88,6	93,4
		K-En Yakın Komşu	Sınama Seti Yaklaşımı	90,47	9,52	98,9	90,1	94,3
			K-Katlı Çapraz Doğ.	90,85	9,15	99,2	90,3	94,5
	Naive Bayes	Ortalama	Sınama Seti Yaklaşımı	90,47	9,52	95,5	93,4	94,4

Tablo 3.8.(Devam) Yöntemlerin değerledirme istatistikleri

			K-Katlı Çapraz Doğ.	89,54	10,45	94,3	92,7	93,5
		K-En Yakın Komşu	Sınama Seti Yaklaşımı	**91,42**	8,57	92	97,6	94,7
			K-Katlı Çapraz Doğ.	87,58	12,41	92,2	92,2	92,2
Dönem 2	**Karar Ağaçalrı**	Ortalama	Sınama Seti Yaklaşımı	**93,33**	6,66	96,2	96,2	96,2
			K-Katlı Çapraz Doğ.	90,63	9,36	94,9	94,2	94,6
		K-En Yakın Komşu	Sınama Seti Yaklaşımı	**93,33**	6,66	95,1	97,5	96,3
			K-Katlı Çapraz Doğ.	91,63	8,36	95,3	95	95,1
	Yapay Sinir Ağları	Ortalama	Sınama Seti Yaklaşımı	90,19	9,80	94,3	94,3	94,3
			K-Katlı Çapraz Doğ.	91,63	8,36	93,9	96,5	95,2
		K-En Yakın Komşu	Sınama Seti Yaklaşımı	92,15	7,84	96,7	94,6	95,6
			K-Katlı Çapraz Doğ.	**92,30**	7,69	97,3	94	95,6
	Random Forest	Ortalama	Sınama Seti Yaklaşımı	**95,09**	4,90	100	94,4	97,1
			K-Katlı Çapraz Doğ.	89,96	10,03	93,8	94,6	94,2
		K-En Yakın Komşu	Sınama Seti Yaklaşımı	94,11	5,88	97,8	95,7	96,7
			K-Katlı Çapraz Doğ.	92,97	7,02	96,5	95,4	95,9
	K En Yakın Komşu	Ortalama	Sınama Seti Yaklaşımı	88,23	11,76	93,3	93,3	93,3
			K-Katlı Çapraz Doğ.	87,96	12,04	94,6	91,7	93,1
		K-En Yakın Komşu	Sınama Seti Yaklaşımı	**92,15**	7,84	97,7	93,5	95,6
			K-Katlı Çapraz Doğ.	90,3	9,69	96,5	92,5	94,5
	Niave Bayes	Ortalama	Sınama Seti Yaklaşımı	**91,17**	8,82	91	98,8	94,7
			K-Katlı Çapraz Doğ.	89,63	10,36	91,1	96,7	938

Tablo 3.8.(Devam) Yöntemlerin değerledirme istatistikleri

		K-En Yakın Komşu	K-Katlı Çapraz Doğ.	89,96	10,03	89,5	98,7	93,9
Dönem 3	**Karar Ağaçları**	Ortalama	Sınama Seti Yaklaşımı	91,86	8,14	95,1	96,3	95,7
			K-Katlı Çapraz Doğ.	**91,25**	8,74	95,9	94,9	95,4
		K-En Yakın Komşu	Sınama Seti Yaklaşımı	88,77	11,22	97,7	90,5	94
			K-Katlı Çapraz Doğ.	90,21	9,79	94,4	95,1	94,8
	Yapay Sinir Ağları	Ortalama	Sınama Seti Yaklaşımı	88,77	11,22	94,6	93,5	94,1
			K-Katlı Çapraz Doğ.	94,05	5,94	95	98,9	96,9
		K-En Yakın Komşu	Sınama Seti Yaklaşımı	**96,93**	3,06	100	96,9	98,4
			K-Katlı Çapraz Doğ.	94,75	5,24	100	94,7	97,3
	Random Forest	Ortalama	Sınama Seti Yaklaşımı	92,85	7,14	98,9	98,9	96,3
			K-Katlı Çapraz Doğ.	94,05	5,94	94,7	99,3	96,9
		K-En Yakın Komşu	Sınama Seti Yaklaşımı	**95,91**	4,08	97,9	97,9	97,9
			K-Katlı Çapraz Doğ.	94,75	5,24	99,6	95,1	96,3
	K En Yakın Komşu	Ortalama	Sınama Seti Yaklaşımı	90,81	9,18	98,9	91,8	95,2
			K-Katlı Çapraz Doğ.	93,7	6,29	99,3	94,4	96,8
		K-En Yakın Komşu	Sınama Seti Yaklaşımı	92,85	7,14	89,9	93,8	96,3
			K-Katlı Çapraz Doğ.	**94,4**	5,59	100	94,4	97,1
	Niave Bayes	Ortalama	Sınama Seti Yaklaşımı	89,79	10,20	91,4	97,7	94,4
			K-Katlı Çapraz Doğ.	86,36	13,63	87,8	97,5	92,4
		K-En Yakın Komşu	Sınama Seti Yaklaşımı	**91,86**	8,14	92,7	98,7	95,6
			K-Katlı Çapraz Doğ.	86,36	13,63	87,8	98,7	95,6

4. SONUÇLAR VE ÖNERİLER

Veri madenciliği, makine öğrenmesi, istatistik, veri tabanı sistemleri gibi farklı disiplinleri bir araya getirerek, atıl durumdaki büyük veri yığınları içerisinde saklı durumda bulunan anlamlı bilgiye ulaşmayı amaçlayan süreçtir. Bu süreç sonucunda elde edilen anlamlı bilginin, iş ve karar süreçlerine olumlu katkılar sağlaması beklenir. Veri madenciliği kullanım alanları ve önemi her geçen gün artmaktadır. Veri madenciliği yöntemleri, eğitim-öğretim faliyetleri sürecinde oluşan veri kümeleri üzerinde de sıklıkla uygulanmaktadır. Elde edilen anlamlı bilgiden, eğitim-öğretim faliyetleri sürecinde karşılaşılan problemlerin çözümünde ve süreçlerin kalite ve verimliliğinin artırılmasına yönelik çalışmalarda faydalanılmaktadır.

Bu tez çalışmasında, Tıp Fakültesi öğrencilerinin kurul derslerine ait sınav sonuç verileri kullanılmıştır. Uygulamada kullanılan veri seti, bir kamu ünivesitesinde aktif olarak kullanılmakta olan öğrenci bilgi sitemi veri tabanından gerekli izinler alınarak temin edilmiştir. Veri seti, ders kurulu sınavları altında yer alan, 51 adet alt kurul dersine ait sınav sonuçları ve dönem sonu başarı durum verilerinden oluşmaktadır. Hedef değişken olarak öğrencilerin dönem sonu başarı durumları kullanılmıştır. Sınav sonuç verileri üzerinde beş farklı veri madenciliği yöntemi uygulanarak, öğrencilerin akademik başarılarının erken dönemlerde tahmin edilmesinin yanı sıra en başarılı yöntemin belirlenmesi amaçlanmıştır. Çalışmada naive bayes, karar ağaçları, rastgele orman, yapay sinir ağları ve k-en yakın komşuluk yöntemleri kullanılmıştır. Tez çalışmasında CRISP-DM veri madenciliği süreci adımları takip edilmiştir. Çalışma kapsamında eksik veri probleminin çözümüne dönük çalışmalar gerçekleştirilmiştir. Bu çalışmalar kapsamında KNIME veri analiz platformunda yer alan eksik veri probleminin çözümüne dönük yöntemlerin yanı sıra, k en yakın komşu makine öğrenmesi algoritması kullanılmıştır. Ayrıca model doğrulama yöntemi olarak sınama seti yaklaşımı ve k-katlı çapraz doğrulama yöntemleri kullanılmış ve oluşturulan modellerin başarı oranlarına etkisi araştırılmıştır. Oluşturulan veri madenciliği modellerinin performanslarının karşılaştırılması amacıyla performans ölçütlerinden faydalanılmıştır.

Öğrencilere ait akademik başarı durumlarının, eğitim süreçlerinin erken dönemlerinde oluşan veri seti kullanılarak tahmin edilmesi amacı ile oluşturulan veri madenciliği modelleri sonucu elde edilen en yüksek doğruluk oranları, dönem 1 öğrencileri için % 92,38 dönem 2 öğrencileri için % 95,09 ve dönem 3 öğrencileri için % 96,93'tür. Elde edilen doğruluk oranları incelendiğinde, öğrencilere ait akademik başarı durumlarının eğitim süreçlerinin erken dönemlerinde başarılı bir şekilde tahmin edilebileceği saptanmıştır. Eğitim süreçleri sonucu oluşan veri seti ve hazırlanan veri madenciliği modeli ile atıl durumdaki veri seti içesinden anlamlı bilgiye ulaşılabileceği görülmüştür.

Eksik veri probleminin çözümüne dönük olarak kullanılan yöntemlerin doğruluk oranları değerlendirildiğinde, k-en yakın komşu makine öğrenme algoritması kullanılarak oluşturulan modellerin daha başarılı sonuçlar verdiği tespit edilmiştir. Bu durum KNIME veri analiz platformu kullanılarak gerçekleştirilecek çalışmalarda, eksik veri probleminin çözümünde k-en yakın komşu makine öğrenme algoritmasının kullanılabilceğini göstermektedir. KNIME veri analiz platformu kullanılarak gerçekleştirilecek çalışmalarda karşılaşılacak eksik veri problemini için alternatif bir çözüm önerisi sunulmuştur.

Model doğrulama yöntemlerine ait doğruluk oranları incelendiğinde, sınama seti yaklaşımı kullanılarak oluşturulan modellerin, k-katlı çapraz doğrulama yöntemi kullanılarak oluşturulan modellere göre daha yüksek doğruluk oranlarına sahip olduğu görülmüştür. Benzer veri seti kümesi ile yapılacak çalışmalarda kullanılacak model doğrulama yöntemi seçilmesi aşaması için bir fikir oluşturulması sağlanmıştır.

Erken dönemlerde tespit edilen başarısız öğrencilere dönüt sağlanarak akademik başarının artılmasına dönük çalışmalar gerçekleştirilebilir ve akademik başarısızlıkların önüne geçilebilir. Bu sayede akademik başarısızlık oranları düşürülebilir. Ayrıca tez çalışması kapsamında tespit edilen başarısız öğrencilere ait farklı metrikler kullanılarak, başarısız öğrencilerin başarısızlık nedenleri üzerine ayrıntılı analiz çalışmaları gerçekleştirilebilir.

Tıp eğitimi süreçlerinde teorik eğitimlerin yanı sıra pratik ve hasta başı eğitimleri çok önemli bir yer tutmaktadır. Başarılı ve başarısız öğrenci sayılarının erken dönemlerde elde edilmesi, öğrenci sayılarında oluşabilecek ani artış ya da azalışların tespit edilerek,

sonraki dönemler için fiziki mekan ve öğretim üyesi planlamalarının çok daha verimli bir şekilde yapılmasına olanak sağlayacaktır.

Ülkemizde yer alan birçok eğitim kurumunda, tez çalışmasında kullanılan veri setine benzer nitelikte veri kümeleri oluşmaktadır. Geliştirilen veri madenciliği modeli yalnızca uygulamanın gerçekleştirildiği Tıp Fakültesinde değil, benzer veri kümesine sahip tüm eğitim kurumlarında kullanılabilir.

KNIME veri analiz platformunun veri madenciliği uygulamalrında etkin bir biçimde kullanabileceği görülmüştür.

KAYNAKLAR

Abuhanoğlu H., Cankul İ., Ayanoğlu Y., Mezuniyet Öncesi Tıp Eğitimi Maliyetlerinin Belirlenmesi: Tıp Fakültesinde Bir Uygulama, Sağlıkta Performans ve Kalite Dergisi, 2012, **4**(2), 39-65.

Aggarwall C.C., Data Mining: The Textbook, Springer, 3rd ed., New York, 2015.

Aldana W.A., "Data mining industry: emerging trends and new opportunities", Massachusetts Institute of Technology, 2000, **11,** 487-499.

Altay A., Üniversite Kütüphanelerinde Veri Madenciliği Uygulamaları: Kırklareli Üniversitesi Kütüphane ve Dokümantasyon Daire Başkanlığı Örneği, Ağrı İbrahim Çeçen Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 2019, **5**(2), 287-316.

Altun M., Öğrenci Akademik Performansinin Kestirilmesine İlişkin Bir Model Önerisi: Veri Madenciliğine Dayali Bir Çalışma, Doktora Tezi, Akdeniz Üniversitesi, Eğitim Bilimleri Enstitüsü, Antalya, 2019, 602297.

Akçapınar G., Çevrimiçi Öğrenme Ortamındaki Etkileşim Verilerine Göre Öğrencilerin Akademik Performanslarının Veri Madenciliği Yaklaşımı ile Modellenmesi, Doktora Tezi, Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara, 2014, 381422.

Akküçük U., Veri madenciliği kümeleme ve sınıflama algoritmaları, 1. Baskı, Yalın Yayıncılık, İstanbul, 2011.

Akı M.O., Sürücü Uykululuğunun Gerçek Zamanlı Görüntü İşleme ve Makine Öğrenmesi Teknikleri ile Tespitine Yönelik Bir Sistem Tasarımı ve Uygulaması, Doktora Tezi , Trakya Üniversitesi, Fen Bilimleri Enstitüsü, Edirne, 2017, 469084.

Aydemir B., Veri Madenciliği Yöntemleri Kullanarak Meslek Yüksek Okulu Öğrencilerinin Başarı Tahmini, Pamukkale Üniversitesi, Fen Bilimleri Enstitüsü, Denizli, 2017, 486838.

Breiman L., Random Forests, Machine Learning, 2001, **45**, 5-32.

Bırtıl F.S., Kız Meslek Lisesi Öğrencilerinin Akademik Başarısızlık Nedenlerinin Veri Madenciliği Tekniği ile Analizi, Yüksek Lisan Tezi, Afyon Kocatepe Üniversitesi, Fen Bilimleri Enstitüsü, 2011, 283421.

Bilgiç E., R Programlama Dili İle Pazar Sepet Analizi: Muş İl Merkezindeki Bir Süpermarkette Tüketicilerin Satın Alma Davranışlarının Tespiti Üzerine Bir Uygulama. Anemon Muş Alparslan Üniversitesi Sosyal Bilimler Dergisi, 2019, **7**(3), 89-97.

Buluz B., Akademik Başarının Modellenmesinde Çizge Madenciliği Yaklaşımı, Yüksek Lisans Tezi, Gabze Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Kocaeli, 2017, 484547.

Chapman P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R., CRISPDM 1.0 step-by-step data mining guide, The CRISP-DM Consortium, CRISPMWP-1104, 2000.

Chien C.F., Chen L.F., "Data Mining to Improve Personnel Selection and Enhance Human Capital: A Case Study in High-Technology Industry", Expert Systems with Applications, 2008, 34, 280-290.

Demir E., Dinçer S., Üretim Sektöründe Veri Madenciliği Uygulamaları: Literatür Taraması, Anadolu Üniversitesi İşletme Fakültesi Dergisi, 2020, **2**(1), 1-2.

Doğan O., Bir E-Ticaret Sitesi Kullanıcı Hesaplarında Şifre Yapılarının Birliktelik Kuralları ile İncelenmesi, Journal of Internet Applications and Management, 2015, **6**(2), 49-61.

Ekelik H., Altaş D., Dijital Reklam Verilerinden Yararlanarak Potansiyel Konut Alicilarinin Rastgele Orman Yöntemiyle Sınıflandırılması, Journal of Research in Economics, 2019, **3**(1), 28-45.

Göral M.A., Kredi Kartı Başvuru Aşamasında Sahtecilik Tespiti İçin Bir Veri Madenciliği Modeli, Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul, 2007.

Marr B., Büyük Veri İş Başında, 1. Baskı, MediaCat, İstanbul, 2017.

Şengür D., Öğrencilerin Akademik Başarılarının Veri Madenciliği Metotları ile Tahmini, Yüksek Lisans Tezi, Fırat Üniversitesi, Eğitim Bilimleri Enstitüsü, 2013, 333873.

Şeker Ş.E., İş Zekası ve Veri Madenciliği, 1. Baskı, Cinius, İstanbul, 2013.

Şeker Ş.E., CRISP-DM: Endüstriler Arası Standart İşleme – Veri Madenciliği için (Cross Industry Standard Processing – Data Mining), YBS Ansiklopedi, 2018, **5**(1), 10-17.

Savaş S., Topaloğlu N., YılmazM., Veri Madenciliği ve Türkiyedki Uygulama Örnekleri, İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 2012, **3**(1), 1-23.

Nayak S.C., Misra B.B., Behera H.S., "Impact of Data Normalization on Stock Index Forecasting", Department of Computer Science&Engineering VSS University of Technology, 2014, **6**, 257-269.

Piramuthu S., "Evaluating Feature Selection Methods for Learning in Data Mining Applications" European Journal of Operational Research, Article in Press, 2003, 1-11.

Garcia S., Ramírez-Gallego S., Luengo J., Big data preprocessing: methods and prospects, Big Data Analytics, 2016, **9**, 1-9.

Oğuzlar A., Veri Ön İşleme, Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 2003, 21, 67-76.

Özdemir Ş., Eğitimde Veri Madenciliği ve Öğrenci Akademik Başarı Öngörüsüne İlişkin Bir Uygulama, Doktora Tezi, İstanbul Üniversitesi, Fen Bilimleri Enstitüsü, 2016, 418777.

Özekes S., Veri Madenciliği Modelleri ve Uygulama Alanları, İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 2003, **2**(3), 65-82.

Nasuhoğlu H., Eczacılık Sektöründe Yapay Sinir Ağları Ve Zaman Serileri Analizi Ile Talep Tahmini, Yüksek Lisans Tezi, Maltepe Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul, 2019, 611168.

Keskin M.V., Büyük Veride Makine Öğrenmesi Uygulaması, Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul, 2018, 505664.

Kıyak E., CRISP-DM Yöntembilimi Kullanilarak Deniz Kuvvetleri Verisi Üzerinde Veri Madenciliği Sınıflandırma Tekniklerinin Karşılaştırılması, Yüksek Lisans Tezi, Kocaeli Üniversitesi, Fen Bilimleri Enstitüsü, Koaeli, 2006, 197935.

Yılmaz N., İnsan Kaynakları Yönetimi'nin Görünen Yüzü: Fortune 500 İşletmeleri Web İçerik Analizi, Yüksek Lisans Tezi, Namık Kemal Üniversitesi, Sosyal Bilimler Enstitüsü, Tekirdağ, 2015, 414661.

Han J., Pei J., & Kamber M., Data mining: concepts and techniques, 3rd. ed, Waltham: Elsevier, Burlington, 2011.

Ho, T.K., The random subspace method for constructing decision forests, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, **20**(8), 832-844.

TAŞCI, A.E., Onan A., K En Yakın Komşu Algoritması Parametrelerinin Sınıflandırma Performansı Üzerine Etkisinin İncelenmesi, Akademik Bilişim Konferansı, Aydın, 2016.

Tekerek A., Veri Madenciliği Süreçleri ve Açık Kaynak Kodlu Veri Madenciliği Araçları, XIII. Akademik Bilişim Konferansı, İnönü Üniversitesi, Malatya, 2011.

URL-1: https://www.mygreatlearning.com/blog/cross-validation/, (Ziyaret Tarihi: 1 Şubat 2021).

URL-2: https://tr.wikipedia.org/wiki/KNIME, (Ziyaret Tarihi: 1 Şubat 2021).

URL-3: https://www.knime.com/model-visualize, (Ziyaret Tarihi: 1 Şubat 2021).

URL-4: https://www.knime.com/deploy-manage, (Ziyaret Tarihi: 1 Şubat 2021).

URL-5: https://www.veribilimiokulu.com/blog/yapay-sinir-aglari/, (Ziyaret Tarihi: 12 Mart 2021).

URL-6: https://docs.python.org/3/tutorial/index.html, (Ziyaret Tarihi: 12 Mart 2021).

URL-7: https://tr.wikipedia.org/wiki/Python_(programlama_dili), (Ziyaret Tarihi: 12 Mart 2021).

Varol Altay E., Alatas B., Nicel Birliktelik Kural Madenciliği İçin Baskın Olmayan Sıralama Genetik Algoritma-II'nin Duyarlılık Analizi, Bilişim Teknolojileri Dergisi, 2020, **13** (1), 37-46.

Witten I. H., Frank E., & Hall M. A., Data Mining: Practical Machine Learning Tools and Techniques, Second Edition, Elsevier, San Francisco, 2011.

Yakut Ö., EKG İşaretlerindeki Aritmilerin Yumuşak Hesaplama Algoritmaları Kullanılarak Hesaplanması, Doktora Tezi, Kocaeli Üniversitesi, Fen Bilimleri Enstitüsü,Kocaeli, 2018, 536633.

Yıldıran A., Kandemir S., Yağış Miktarının Yapay Sinir Ağları ile Tahmini, Bilecik Şeyh Edebali Üniversitesi Fen Bilimleri Dergisi, 2018, **5**(2), 97-104.

EKLER

Ek-A

Dönem 1 Database Reader düğümü SQL sorgusu

```
SELECT * FROM (
SELECT DISTINCT
dbo.sinavSonuc.OgrenciID, dbo.sinavSonuc.Net, dbo.altKurul.altKurulAd
FROM dbo.sinavSonuc INNER JOIN
dbo.altKurul ON dbo.sinavSonuc.AltKurulId = dbo.altKurul.altKurulId
INNER JOIN
dbo.donemBasariDurum ON dbo.sinavSonuc.OgrenciID =
dbo.donemBasariDurum.ogrenciId
WHERE   (dbo.sinavSonuc.Sinif = 1) AND
(dbo.donemBasariDurum.egitimYili = 20162017)
) StudentResults
PIVOT (
  SUM([Net])
  FOR [altKurulAd]
  IN (
[Halk Saglıgı-1],
[Ruh Saglıgı ve Hastalıkları-1],
[Tıbbi Biyokimya-1],
[Tıbbi Biyoloji-1],
[Tıp Tarihi ve Deontoloji-1],
[Aile Hekimligi-2],
[Biyoistatistik ve Tıp Bilisimi-2],
[Fizyoloji-2],
[Radyoloji-2],
[Ruh Saglıgı ve Hastalıkları-2],
[Tıbbi Biyokimya-2],
[Tıbbi Biyoloji-2],
[Tıp Tarihi ve Deontoloji-2],
[Acil Tıp-3],
[Aile Hekimligi-3],
[Biyoistatistik ve Tıp Bilisimi-3],
[Histoloji ve Embriyoloji-3],
[Radyoloji-3],
[Ruh Saglıgı ve Hastalıkları-3],
[Tıbbi Biyokimya-3],
[Tıbbi Biyoloji-3]
  )
) AS PivotTable
```

Ek-B

Dönem 2 Database Reader düğümü SQL sorgusu

```
SELECT * FROM (
SELECT DISTINCT
dbo.sinavSonuc.OgrenciID, dbo.sinavSonuc.Net, dbo.altKurul.altKurulAd
FROM           dbo.sinavSonuc INNER JOIN
dbo.altKurul ON dbo.sinavSonuc.AltKurulId = dbo.altKurul.altKurulId
INNER JOIN
dbo.donemBasariDurum ON dbo.sinavSonuc.OgrenciID =
dbo.donemBasariDurum.ogrenciId
WHERE       (dbo.sinavSonuc.Sinif = 2) AND
(dbo.donemBasariDurum.egitimYili = 20172018)
) StudentResults
PIVOT (
SUM([Net])
FOR [altKurulAd]
IN (
[Aile Hekimligi-1],
[Anatomi-1],
[Fizyoloji-1],
[Histoloji ve Embriyoloji-1],
[Kardiyoloji-1],
[Tıbbi Biyokimya-1],
[Anatomi-2],
[Fizyoloji-2],
[Histoloji ve Embriyoloji-2],
[Tıbbi Biyokimya-2],
[Anatomi-3],
[Fizyoloji-3],
[Histoloji ve Embriyoloji-3],
[Tıbbi Biyokimya-3])
) AS PivotTable
```

Ek-C

Dönem 1 Database Reader düğümü SQL sorgusu

```
SELECT * FROM (
SELECT DISTINCT
dbo.sinavSonuc.OgrenciID, dbo.sinavSonuc.Net, dbo.altKurul.altKurulAd
FROM           dbo.sinavSonuc INNER JOIN
dbo.altKurul ON dbo.sinavSonuc.AltKurulId = dbo.altKurul.altKurulId
INNER JOIN
dbo.donemBasariDurum ON dbo.sinavSonuc.OgrenciID =
dbo.donemBasariDurum.ogrenciId
WHERE        (dbo.sinavSonuc.Sinif = 3) AND
(dbo.donemBasariDurum.egitimYili = 20182019)
) StudentResults
PIVOT (
SUM([Net])
FOR [altKurulAd]
IN (
[Enfeksiyon Hastalıkları ve Klinik Mikrobiyoloji-1],
[Radyasyon Onkolojisi-1],
[Tıbbi Farmakoloji-1],
[Tıbbi Genetik-1],
[Tıbbi Mikrobiyoloji-1],
[Tıbbi Patoloji-1],
[Cocuk Saglıgı ve Hastalıkları-2],
[Gögüs Cerrahisi-2],
[Gögüs Hastalıkları-2],
[Kardiyoloji-2],
[Tıbbi Farmakoloji-2],
[Tıbbi Mikrobiyoloji-2],
[Tıbbi Patoloji-2],
[Cocuk Saglıgı ve Hastalıkları-3],
[Enfeksiyon Hastalıkları ve Klinik Mikrobiyoloji-3],
[Genel Cerrahi-3],
[ic Hastalıkları-3],
[Tıbbi Farmakoloji-3],
[Tıbbi Mikrobiyoloji-3
[Tıbbi Patoloji-3],
[Cocuk Saglıgı ve Hastalıkları-4],
[ic Hastalıkları-4],
[Tıbbi Farmakoloji-4],
[Tıbbi Mikrobiyoloji-4],
[Tıbbi Patoloji-4],
[Üroloji-4]
  )
) AS PivotTable
```

Ek-D

Python Script düğümü eksik veri oluşturma kodu

```
# Copy input to output
#output_table = input_table
import pandas as pd
import numpy as np
import matplotlib as plt
from sklearn import preprocessing
import random
deger=random.randint(5,10)
df =input_table
deger2=len(df.index) #satır sayısı
deger3=round((deger2*10)/100)
output_table = df
for i in range(0,14):
  for j in range(0,int(deger3)):
          deger4=random.randint(1,309)
          df.iloc[int(deger4),i]=float('NaN')
```

Ek-E

K-en yakın komşu makine öğrenme algoritması kodu

```
# Copy input to output
#output_table = input_table
import pandas as pd
import numpy as np
from sklearn import preprocessing
df =input_table
from sklearn.impute import KNNImputer
from sklearn.impute import KNNImputer
imputer = KNNImputer(n_neighbors=3)
imputed = imputer.fit_transform(df)
df_imputed = pd.DataFrame(imputed, columns=df.columns)
output_table = df_imputed
```

KİŞİSEL YAYIN VE ESERLER

Maden E., Yıldırım M., Diri S., Kocaeli Mühendislik Fakültesi Mezunlarının Akademik Başarılarının Veri Madenciliği Yöntemleri Kullanılarak İncelenmesi, *International Congress Of Academic Research*, Bolu, Türkiye, 17-19 Şubat 2020.

ÖZGEÇMİŞ

İlk orta ve lise öğrenimini Ankarada tamamladı. 2011 yılında girdiği Kocaeli Üniversitesi Bilgisayar Mühendisliği Bölümü'nden 2016 yılında mezun oldu. 2017 yılında Kocaeli Üniversitesi Bilişim Sistemleri Mühendisliği bölümünde başladığı yüksek lisans eğitimine devam etmektedir. Halen Hacettepe Üniversitesi Tıp Fakültesi bilgi işlem biriminde görev yapmaktadır.