T.C.
İSTANBUL AYDIN ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

MAKİNE ÖĞRENMESİ ALGORİTMALARIYLA KALP HASTALIKLARININ TESPİT EDİLMESİNE YÖNELİK PERFORMANS ANALİZİ

YÜKSEK LİSANS TEZİ

Elif ÇİL

Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Programı

AĞUSTOS, 2022

T.C.
İSTANBUL AYDIN ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

MAKİNE ÖĞRENMESİ ALGORİTMALARIYLA KALP HASTALIKLARININ TESPİT EDİLMESİNE YÖNELİK PERFORMANS ANALİZİ

YÜKSEK LİSANS TEZİ

Elif ÇİL
(Y1813.010043)

Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Ali GÜNEŞ

AĞUSTOS, 2022

ONAY FORMU

ONUR SÖZÜ

Yüksek Lisans tezi olarak sunduğum “Makine Öğrenmesi Algoritmalarıyla Kalp Hastalıklarının Tespit Edilmesine Yönelik Performans Analizi” adlı çalışmanın, tezin proje safhasından sonuçlanmasına kadarki bütün süreçlerde bilimsel ahlak ve geleneklere aykırı düşecek bir yardıma başvurulmaksızın yazıldığını ve yararlandığım eserlerin kaynakçada gösterilenlerden oluştuğunu, bunlara atıf yapılarak yararlanılmış olduğunu belirtir ve onurumla beyan ederim (30/06/2022).

Elif ÇİL

ÖNSÖZ

Tez çalışmamda planlanmasında, araştırılmasında, yürütülmesinde ve oluşumunda ilgi ve desteğini esirgemeyen, engin bilgi ve tecrübelerinden yararlandığım, yönlendirme ve bilgilendirmeleriyle çalışmamı bilimsel temeller ışığında şekillendiren sayın hocam Prof. Dr. Ali GÜNEŞ'e ve manevi desteğini esirgemeyen Hülya ÇİL'e sonsuz teşekkürlerimi sunarım.

AĞUSTOS, 2022 Elif ÇİL

MAKİNE ÖĞRENMESİ ALGORİTMALARIYLA KALP HASTALIKLARININ TESPİT EDİLMESİNE YÖNELİK PERFORMANS ANALİZİ

ÖZET

İnsan yaşamında kişiler birçok rahatsızlıkla karşılaşmaktadırlar. Hastalık bireylerin yaşamlarında büyük bir yere sahiptir. Dünya ve Türkiye'de hastalıklar içerisinde kalp rahatsızlıkları ciddi bir yer tutmaktadır. Kalp ile hastalıkların meydana gelmesinde birden fazla sebep bulunmakta ve kişiden kişiye etkileri farklılık göstermektedir. Fakat bireylerin kalp rahatsızlığı sebebi ile hayatının sona ermesi durumunu engelleyebilmek adına birçok tedavi yöntemi geliştirilmiştir. Kalp hastalıklarından erken teşhis ile tedavinin önemi çok büyük yer edinmektedir. Yaşam içerisinde bu yöntem ve teknikleri sürekli gelişim ve değişim göstermektedir. Makine öğrenme yöntemleri günümüzde pek çok alanda çok aktif bir şekilde kullanım göstermektedir. Ancak sağlık alanında makine yöntemlerinin kullanımı yeni yaygınlık göstermektedir. Bu sebep ile bu çalışmada içerisinde makine öğrenme yöntemleri kullanılacaktır. Kalp hastalıların tespit edilebilmesinde makine öğrenme yöntemlerinden hangisinin daha başarılı olduğunu bir sonuç vereceğini incelenmiştir.

Anahtar Kelimeler: Makine Ögrenmesi, Yapay Zeka, Veri İşleme Teknikleri, Kalp Hastalıkları

PERFORMANCE ANALYSİS FOR DETECTİNG HEART DİSEASES WİTH MACHİNE LEARNİNG ALGORİTHMS

ABSTRACT

In human life, people face many ailments. The disease has a great place in the lives of individuals. Heart diseases have a serious place among diseases in the world and in Turkey. There are more than one reason for the occurrence of heart diseases and their effects differ from person to person. However, many treatment methods have been developed in order to prevent the death of individuals due to heart disease. The importance of diagnosis and treatment of heart diseases is very important. These methods and techniques show continuous development and change in life. Machine learning methods are used very actively in many fields today. However, the use of machine methods in the field of health shows a new prevalence. For this reason, machine learning methods will be used in this study. It has been examined that which of the machine learning methods is more successful in detecting heart diseases will give a result.

Keywords: Machine Learning, Artificial Intelligence, Data Processing Techniques, Heart Diseases.

İÇİNDEKİLER

ŞEKİLLER LİSTESİ

ÇİZELGELER LİSTESİ

KISALTMALAR LİSTESİ

AA : Aort Aterosklerozu

BRFSS : The Behavioral Risk Factor Surveillance System

CDSS : Clinical Decision Support System

ÇKAYSA : Çok Katmanlı Algılayıcı Yapay Sinir Ağları

EKG : Elektrokardiyografi

EM : Beklenti Maksimizasyonu

EuroSCORE : The European System for Cardiac Operative Risk Assessment

KA : Karar Ağaçları

KAH : Koroner Arter Hastalığı

KKH : Koroner Kalp Hastalığı

KNN : K-En Yakın Komşu

LR : Logistik Regresyon

NB : Naive Bayes

PAH : Periferik Arter Hastalığı

SVH : Serebrovasküler Hastalık

SVM : Destek Vektör Makineleri

TEKHARF : Türk Erişkinlerinde Kalp Hastalıkları ve Risk Faktörleri

TOA : Topluluk Öğrenme Algoritmaları

TÜİK : Türkiye İstatistik Kurumu

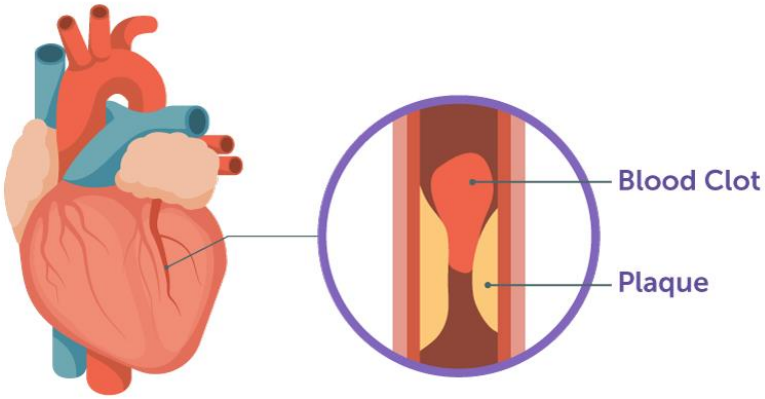
YSA : Yapay Sinir Ağları

I. GİRİŞ

Kalp hastalıkları modern dünyada her geçen gün artmaktadır. Dünya Sağlık Örgütü'ne göre, yılda 17,9 milyon insan kalp krizi ve felç gibi kalp hastalıklarından ölmektedir. Dünya genelinde kalp krizi ve felç sebebiyle meydana gelen ölümlerin üçte biri 70 yaşın altındaki kişilerde daha erken meydana gelmektedir (URL-1). Kalp hastalıkları kardiyovasküler hastalıklar olarak adlandırılır. Koroner kalp hastalığı, serebrovasküler hastalık, romatizmal kalp hastalığı ve diğer durumları içermektedir. Kalp hastalıklarından kaynaklanan beş ölümün dördünden fazlası kalp krizi ve felçten kaynaklanmaktadır (URL-1). Kalp hastalıkları Türkiye'de de can kaybına neden olan önemli ölümcül hastalıklardan biridir. TÜİK ölüm nedeni istatistiklerinde ülkemizde kalp hastalıkları sebebiyle gerçekleşen ölüm oranı %36.8 olarak belirtilmektedir (URL-2). Türk Erişkinlerinde Kalp Hastalıkları ve Risk Faktörleri (TEKHARF) çalışmasında ülkemizde elde edilen 2009 yılı sonuçlarına göre, 45-74 yaş aralığında olan koroner damarların tıkanması ve daralması olarak nitelendirilen kalp hastalığı kaynaklı ölümlerin kadınlarda 3.84, erkeklerde 7.64 oranında olduğu raporlanmıştır. Türkiye'nin elde edilen verilere göre kalp hastalıkları bakımından 30 Avrupa ülkesine göre en yüksek seviyede olduğu belirlenmiştir (Onat vd., 2009).

Kalp hastalıkları, kalp ve kan damarlarının oluşturduğu bir grup bozukluğudur. Kalp hastalığı terimi, kalbi etkileyen çeşitli hastalıkları kapsar. Tamamı kardiyovasküler hastalık olarak adlandırılan kalp hastalıkları; serebrovasküler hastalık, koroner kalp hastalığı, romatizmal kalp hastalığı ve diğer durumları içerir.

Kalp hastalığı teşhisi konmadan önce farklı testler yapılmaktadır. Bunlardan önde gelenleri oskültasyon, EKG, tansiyon, kolesterol ve kan şekeri testleridir. Bu testler genellikle kapsamlıdır (Okcu, 2011).



Şekil 1 Kalp ve Kan Damarları

Kardiyovasküler hastalıkların birkaç farklı semptomla ilişkili olması hastalığı hızlı teşhis etmeyi zorlaştırmaktadır. Teşhisin geç yapılması tedavi sürecini önemli ölçüde etkilemekte, tedavi edilebilir düzeyde iken tanı konulamadığı takdirde hastalar yaşam süreci boyunca ilaç kullanmak zorunda kalmaktadır. Bu yüzyılın en önemli bilimsel misyonlarından biri, bilgisayar bilimlerinin tıpla bütünleştirilmesidir. Teknoloji ve beraberinde getirdiği yenilikler yaşam kalitesini her zaman olumlu etkilemektedir. Son zamanlarda keşfedilen yeni teknik ve yöntemler teknoloji ile bileştirilerek tedavi ve tanı süreçlerinin, hastalarla iletişimin, sağlık hizmetleriyle ile ilgili süreçlerin ve sağlık kurumlarının yönetimsel süreçlerinin verimli bir şekilde yürütülmesini ve optimize edilmesini sağlamaktadır. Günümüzde hastaneler sağlık veya hasta verilerini yönetmek amacıyla hastane bilgi sistemleri kullanmaktadır. Bu sistemler çok büyük miktarda veri üretir. Tıbbi veri kümeleri geniş çapta dağılmış, heterojen ve çok büyüktür. Bu veri setlerinin organize edilmesi ve hastane yönetim sistemleri ile entegre edilmesi gerekmektedir. Hastanelerde elde edilen büyük veri klinik durumu desteklemek için nadiren kullanılmaktadır. Kalp hastalığının teşhisi için var olan verilerin sınıflandırılması karmaşık soruları yanıtlama imkanı sunmaktadır. Bu durum sağlık uzmanlarına karar verme süreçlerinde olumlu değişimler getirmekle birlikte bilgi kaynağı olarak kullanılabilmektedir. Tıbbi verilerin ifadesine dayalı olarak kalp hastalığının teşhis etmek için kullanılan iki temel gerçekçi yaklaşım istatistik ve makine öğrenimidir. Karar destek sistemleri, klinik testlerin daha düşük bir maliyetle gerçekleştirilmesine ve hastalık tanı sürecinin minimuma indirilmesine yardımcı olmaktadır.

Bu çalışma, kalp hastalığının teşhis edilmesinde kullanılan büyük veriyi veri işleme tekniklerini kullanarak optimize etmeyi, işlenen veriye makine öğrenme tekniklerini uygulamayı ve algoritmaların performanslarını analiz etmeyi amaçlamaktadır. Kalp hastalıklarının tespit edilebilmesi için beş adet makine öğrenmesi algoritması özniteliklere uygulanmış, algoritmaların başarı oranları tespit edilerek sonuçlar karşılaştırılmış ve en iyi sonucu veren algoritma ile kalp hastalığı teşhisi konulması amaçlanmıştır.

Çalışmada işlenen veri seti kalp hastalığının ikili sınıflandırması için kullanılmaktadır. Kalp hastalığı tanısı olanlar 1, olmayanlar 2 olarak sınıflandırılmıştır. BRFSS veri seti 253680 anket yanıtı içermektedir. 229787 katılımcının kalp hastalığı tanısı bulunmamaktadır. 23893 kişinin ise kalp hastalığı tanısı vardır. BRFSS veri seti ham olarak 330 sütuna sahiptir ancak kalp hastalığını etkileyen faktörlere ilişkin bilimsel araştırmalarına dayanarak bu analize yalnızca belirli özellikler dahil edilmiştir.

Tez beş bölümden oluşmaktadır. Tezin birinci bölümünde kalp hastalıkları, nedenleri, etki eden faktörler, belirtileri ve çeşitleri açıklanmıştır. İkinci bölümünde yöntem ve teknikler, makine öğrenmesine genel bakış, makine öğrenmesi algoritmaları, sağlık alanında kullanılan makine öğrenme tekniklerinden bahsedilmiştir. Üçüncü bölümünde veri, veri işleme teknikleri, veri temizleme ve ölçekleme konusu ele alınmıştır. Dördüncü bölümünde weka ile yapay zeka algoritmaları kullanılarak veri işleme, algoritmaların sonuçları gösterilmiş ve değerlendirmesi yapılmıştır. Beşinci ve son bölümde ise sonuçlar ve sonuçların analizi yapılmıştır.

II. LİTERATÜR TARAMASI

Yapay zeka yöntemleri kullanılarak sağlık alanında hastalıkların tahminin konu alan pek çok çalışma yapılmıştır. Makine öğrenmesi algoritmaları kullanılarak hastalıkların teşhis edilmesi klinik süreçleri hızlandırmaktadır. Tıp alanında makine öğrenmesi algoritmaları kullanılarak kalp hastalıklarının teşhisi alanında yapılan çalışmalar da bulunmaktadır.

2014 yılında Uysal, Bilen ve Ulukuş tarafından yapılan Twoing algoritması ile kalp hastalığı uygulamasında büyük damarlar özniteliği en ayırt edici özellik olarak bulunmuştur. 2016 yılında Bulut tarafından Adaboost ile kalp krizi risk tespiti çalışması yapılmış, veri setinde bulunan 47 öznitelikten yüksek tansiyonun kalp krizinin tahmin edilmesinde %87.89 başarı sağladığı tespit edilmiştir. 2018 yılında Özmen, Khdr ve Avcı tarafından sınıflandırıcıların kalp hastalığı verileri üzerine performans karşılaştırması çalışmasında Destek Vektör Makinelerinin diğer algoritmalara göre %89.4 ile en başarılı olduğu saptanmıştır. 2019 yılında Özcan, Taşar, Tatar ve Yakut tarafından yapılan çalışmada Destek Vektör Makineleri ve Yapay Sinir Ağları algoritmaları kalp hastalıkları tahmininde sırasıyla %91.67 ve %83.33 başarı elde etmiştir. 2020 yılında Göktaş ve Yağanoğlu tarafından yapılan çalışmada kalp krizi riskinin tahmin edilmesinde kullanılan makine öğrenmesi algoritmalarından %83 oranla en başarılı C4.5 Karar Ağacı Algoritması olmuştur. 2020 yılında Görgün tarafından kalp hastalıklarının teşhisinde 10 farklı makine öğrenmesi algoritması kullanılarak en yüksek başarının %90,16 oranında Rastgele Orman algoritması ile sağlandığı ortaya konmuştur. 2020 yılında Gündoğdu tarafından kalp hastalık risk tahmini çalışmasında Python ile kullanılan özniteliklerden göğüs ağrısı türü'nün %13.43 ile sınıflandırmaya en fazla etkisi olduğu, Random Forest'ın kullanılan sınıflandırıcılar arasında %90.2 oranla en çok başarı sağladığı saptanmıştır. 2021 yılında Coşar ve Deniz tarafından kalp hastalıklarının tahmininde üç farklı sınıflandırma algoritması kullanılarak %88 oranla en başarılı sonuçlar Random Forest algoritması ile elde edilmiştir.

Yılmaz ile Sümer tarafından 2021 senesinde "*Relief Özellik Seçim Yöntem Tabanlı Önerilen Hibrit Model ile Kalp Hastalığı Teşhisi*" isimli bir çalışma gerçekleştirilmiştir. Kalp ritmi problemleri kan damar hastalıkları ile bunun yanı sıra doğuştan gelen kalp sorunları insan yaşamı için ciddi seviyede risk bulunduran kalp hastalıkları başlığı altında bulunmaktadır. Bu çalışma içerisinde kalp hastalıklarının göğüs ağrısı, cinsiyet, kolesterol gibi nitelikler incelenerek çok sık kullanılan makine öğrenmesi yöntemleri olan karar ağaçları (KA), logistik regresyon (LR), K-en yakın komşu (KNN), naive bayes (NB), çok katmanlı yapay sinir ağları (YSA), destek vektör makineleri (SVM) ve Relief özellik çıkarım yöntem tabanlı hibrit bir yöntem tavsiye edilerek kalp hastalıkları için analiz gerçekleştirilmiştir. Yapılmış olan değerlendirmenin neticesinde başka makine öğrenme tekniklerine göre hibrit modelin performansı hem zamansal bakımından hem de doğruluk bakımından çok daha başarılı netice vermiştir.

Bulut 2016 senesinde "*Torbalama sınıflandırıcı kullanarak kalp krizi riski tespiti*" adlı çalışma yapmıştır. Dünyada kardiyovasküler hastalıklar en sık ölüm sebebi olmaktadır. Çalışmada, toplu bir Makine Öğrenimi sınıflandırma algoritması olan Torbalama Yöntemini kullanarak bireylerin kalp krizi riskini tahmin etmek amaçlanmıştır. Kalp krizi geçiren hastalara resmi izinler alındıktan sonra anketler uygulanmıştır. Bu sayede sınıflandırma algoritmalarında kullanılmak üzere önceden tanımlanmış bir veri seti oluşturulmuştur. Uygulamalarda güçlü topluluk sınıflandırıcıları kullanılarak bir birey için kalp krizi riski tespit edilebilmektedir. Ayrıca çapraz doğrulama sürecinde önerilen model regresyonda yüksek performans göstermektedir. Bu nedenle önerilen bu Klinik Karar Destek Sistemi (CDSS), kalp krizi öncesinde bazı önlemlerin alınmasını sağlamaktadır.

Akgül ve arkadaşları 2020 senesinde "*Hipotiroidi Hastalığı Teşhisinde Sınıflandırma Algoritmalarının Kullanımı*" adlı çalışma gerçekleştirmişlerdir. Çalışma tanı sürecinde olan hasta kişilere sorulan soru ve uygulanan test neticeleri kullanarak hipotiroidi hastalığının doğru teşhis seviyesinde artış gösterecek veri madenciliği temeli olan bir sistem ortaya koymaktadır. Diğer amacıysa dolaylı şekilde teşhis için kullanılan girişimsel testlerden ortaya çıkacak komplikasyonlarda azaltma göstermektedir. Bu amaçların doğrultusunda UCI makine öğrenmesi veri tabanında yer alan ve 151 tanesi hipotiroidi geri kalanı hipotiroidi olmayan toplam 3163 örnekten meydana gelen veri seti kullanılarak yeni örneklerin hipotiroidi olup olmadığı ön

görülmektedir. Veri setinde olan dengesiz dağılımı yok etmek amacıyla veri setine değişik örnekleme teknikleri yaparak Lojistik Regresyon, K En Yakın Komşu ve Destek Vektör Makinesi sınıflandırıcılarıyla hipotiroidi hastalığını öğrenilecek modeller meydana getirilmiştir. Yapılan modellerin içerisinde en yüksek performansı, aşırı örnekleme yöntemleri uygulanan veri setiyle eğitilen Lojistik Regresyon sınıflandırıcısı göstermiştir.

Bilgin 2021 senesinde "*Makine Öğrenmesi Algoritmaları Kullanarak Erken Dönemde Diyabet Hastalığı Riskinin Araştırılması*" isimli bir çalışma yapmıştır. Diyabet, insanın yaşam kalitesine önemli ölçüde etki eden, dünyada ve Türkiye'de görülme sıklığı gitgide artan ciddi bir hastalıktır. Özel olarak böbreklere, sinir sistemine, gözlere, kalbe, kan damarları ile uzuvlara zarar vermekte ve ciddi kayıplara sebep olabilmektedir. Bu sebep ile diyabetin önlenmesi ya da zararının en aza indirilebilmesi için erken teşhis ile takip büyük ciddiyet taşımaktadır. Makine öğrenmesi algoritmalarıyla ortaya konan sınıflandırma teknikleri, araştırmacılar tarafından hastalığın risk tahmin modeli için önemli kabul edilmiştir. Çalışmada, diyabet geliştirme ihtimalini ön görmek için 520 denekten alınan bilgilerle oluşturulan bir veri tabanı kullanılmış. Çalışma içerisinde makine öğrenmesi yöntemleri olarak Destek Vektör Makineleri (SVM), Çok Katmanlı Algılayıcı Yapay Sinir Ağları (ÇKAYSA), Topluluk Öğrenme Algoritmaları (TOA), Karar Ağaçları (KA), k-NN Yöntemleri, Doğrusal Ayrım Analizi (DAA) kullanılmıştır. Bu yöntemler içerisinde en yüksek doğruluğu k-NN algoritması vermiş ve bu algoritmayla %99,81 doğruluğa ulaşılmıştır. Çalışma kapsamında geliştirilen bir bilgisayar kullanıcı ara yüzüne en yüksek doğruluk değerini sağlayan algoritma dahil edilerek diyabet erken tanı kitinin gelişimi gerçekleştirilmiştir.

Kaya 2017 senesinde "*Makine öğrenmesi teknikleri ile aritmi tespiti ve yeni öznitelikler ile başarımın artırılması*" adlı çalışma gerçekleştirmiştir. Çalışma içerisinde ilk olarak işarete dair zaman serileri kullanılmış ve EKK vurusunun gruplandırılması yapılmıştır. Bir vuruluk işaretin zaman serisine ilave şekilde türlü boyut indirgeme algoritmalarının performansına olan etkisi incelenmiştir. İlave şekilde çalışma kapsamı daha geniş aritmi çeşitlerini sınıflandıracak biçimde genişletilmiş ve testler yapılmıştır. Bu aritmilerin gruplandırılması amacıyla bir vuruluk işarete dair yeni öznitelikler ortaya konulmuştur. Boyut indirgeme algoritmaları kullanılarak öznitelikler daha küçük boyutlara düşürülmüştür. Deneyler, yapay sinir ağları, k-en

yakın komşu algoritması, destek vektör makinesi ile karar ağaçları sınıflandırıcıları kullanılarak yapılmıştır. Bulgular, duyarlılık, doğruluk, özgünlük, kesinlik ve çalışma süreleri açısından incelenmiştir. Çalışma içerisinde gerçekleştirilen testlerde kullanılmış olan veriler bu alanda standart hale gelen olan MIT-BIH aritmi veri tabanından alınmıştır.

Oğuztürk 2018 senesinde "*Diyabet hastalığının makine öğrenmesi algoritmaları ile en iyi doğru tahmininin elde edilmesi*" isimli bir çalışma yapmıştır. Makine öğrenimi algoritmalarına dayalı erken diyabet teşhisi için hızlı, kolay ve hassas bir tahmin aracı geliştirmek gerekli olmaktadır. Çalışmada kullanılmış olan veri seti, Türkiye'deki diyabetik olmayan ve diyabetli hasta kişilerin sağlık profillerinden meydana gelmektedir. Hasta kişilerin on değişik niteliği giriş değişkeni şeklinde seçilmiş olup, netice değişkeni şeklinde de hasta olup olmadığına ilişkin değerler ele alınmıştır. Diyabetik durumun ön görülmesi için elde edilen veriler, yedi değişik makine öğrenmesi algoritması uygulanarak işleme konulmuştur. Toplam 2657 tane denekten 1860 tanesi algoritmanın eğitimi amacıyla kullanılmış, kalan 797 veri tanesiyse algoritmanın test edilmesi amacıyla ayrı tutulmuştur. Diyabet öngörme modelinin geliştirilebilmesi amacıyla açık kaynak kodlu Orange programı çalışmada kullanılmıştır. Doğruluğunu algoritmanın optimize edebilmek amacıyla değişik kombinasyonlar, beklenti maksimizasyonu (EM) ve gizli düğüm sayısı iterasyonları uygulanmıştır. Yapay sinir ağı algoritmasının, %97,2'lik doğru ön görme başarısı ile en iyi başarıyı ulaştığı belirlenmiştir.

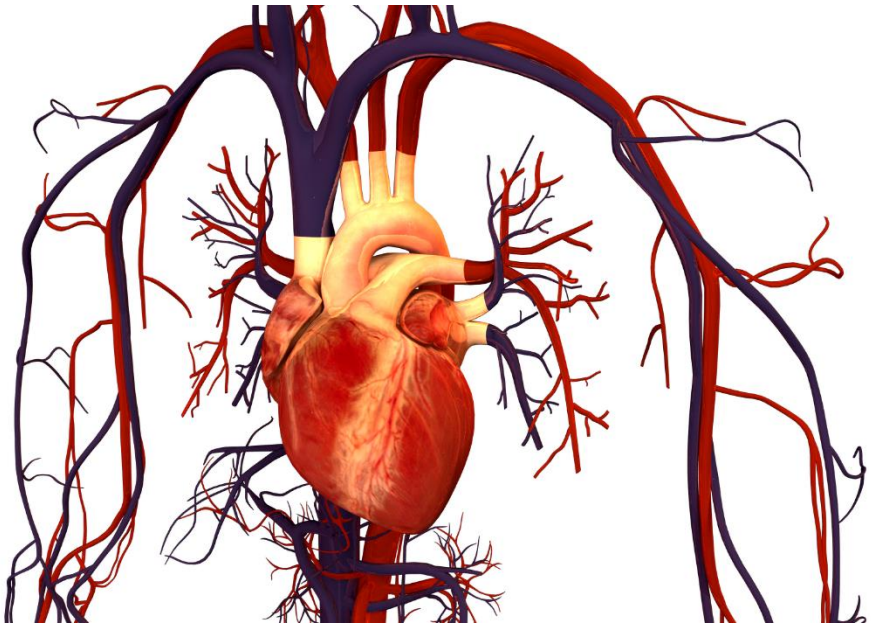
Kartal 2015 senesinde "*Sınıflandırmaya Dayalı Makine Öğrenmesi Teknikleri ve Kardiyolojik Risk Değerlendirmesine İlişkin Bir Uygulama*" isimli çalışma yapmıştır. Araştırmada kullanılan veri seti Acıbadem Maslak Hastanesi'nden elde edilmiştir. Literatürde hastanın kalp cerrahisi sırasında veya kısa bir süre sonra ölüm riskini tahmin etmek için kullanılan EuroSCORE (The European System for Cardiac Operative Risk Assessment) risk faktörleri kullanılmıştır. EuroSCORE'da olduğu gibi veri setindeki gözlemlerde 30 günlük takip bilgisi bulunmadığından ilk olarak hastaların Standart EuroSCORE puanları hesaplanmıştır. Daha sonra bu risk grupları sınıf etiketleri olarak kullanılmış ve tahminler yapılmıştır. Naive Bayes Sınıflandırıcı, Lojistik Regresyon Analizi, k-En Yakın Komşu Algoritması, ID3 ve C4.5 Karar Ağacı Algoritmaları kullanılarak değişik modeller ortaya konuşmuştur. Modellerin performansları kıyaslanmıştır. Veri analizleri R dilinde yazılan kodlarla yapılmıştır.

RStudio, R kodları geliştirmek için bir araç olarak kullanıldı. Lojistik Regresyon Analizinden elde edilen modeller, web üzerinde Shiny (shinyapps.io) aracılığıyla halka açık hale getirilmiştir. En iyi performans gösteren C4.5 karar ağacı modeli için başka bir Shiny uygulaması geliştirilmiştir.

III. TEMEL KAVRAMLAR

A. Kalp Hastalıkları

Kalp hastalıkları günümüzde ölüm nedenlerinde ilk sırada gelmektedir. Özellikle Türkiye gibi gelişmekte olan ülkelerde bu durum daha net bir biçimde görülmektedir. Kalp hastalıklarının gelişme riskinin tahmin edilmesi yetişkin yaşa gelmiş olan bireylerde koruyucu önlemler ve tedavi açısından çok önemlidir (Kültürsay, 2011). Kardiyovasküler sistem kalp ve onun kan damarlarından oluşmaktadır. Kardiyovasküler sistemde meydana gelen bozukluklara kalp hastalığı ya da kardiyovasküler hastalık adı verilmektedir (Farlet vd., 2012).



Şekil 2 Kardiyovasküler Sistem

Kalp sağlığında dolaşım ve kasların önemi büyüktür. Dolaşımın amacı, dokuları besleyebilen ve dokular tarafından sürdürülen enerji kaybını yerine koyabilen sabit bir madde akımının sağlanmasıdır. Kan ve dokular arasındaki ürün alışverişini kolaylaştırmak için kan kılcal damarlardan geçmektedir. Kalp kası, dolaşımı sağlayan gücü sağlamaktadır. Normal durumda, dolaşım mekanizmasında tüm parçalar kalbin çalışmasını kolaylaştırmak ve dolaşımın amacına ulaşmak için birleşmektedir. Bu

ayarın herhangi bir şekilde bozulması, kalbin normal arter basıncını korumaktan utanması anlamına geldiği için, kalp kası üzerinde derhal daha fazla çalışma gerektirmelidir. Kalp bu engeli aşabildiği ve dolaşımı normal bir şekilde sürdürebildiği sürece hiçbir belirti ortaya çıkmaz. Ancak kalp artık dolaşımı verimli bir şekilde sürdüremezse bazı fenomenler ortaya çıkmaktadır (Mackenzie, 2005).

Kalp hastalıkları ya da kardiyovasküler hastalıklar; periferik damar hastalıkları (klaudikasyon veya ekstremite iskemisi), konjenital kalp hastalıkları, konjestif kalp yetmezliği, hipertansif hastalıklar, serebrovasküler hastalık (SVH)'lar (inme), koroner kalp hastalıkları (anjina pektoris, miyokard infarktüsü), romatizmal kalp hastalıkları ve aritmiler gibi tüm kalp ve damar hastalıklarını içine almaktadır (Diehm vd., 2006).

B. Kalp Hastalıklarının Nedenleri

Kalp hastalıklarının nedenleri genetik ve çevresel faktörler olarak ayrılmaktadır. En önemli davranışsal risk faktörleri sağlıksız beslenme, fiziksel hareketsizlik, tütün kullanımı ve zararlı alkol kullanımıdır. Bu risk faktörlerini yoksulluk, stres ve kalıtsal faktörler takip etmektedir (URL-3). Türkiye'deki kardiyovasküler hastalıkların ana faktörleri hipertansiyon, hiperlipidemi, diyabet ve sigara olarak sıralanmaktadır (Abacı, 2011). Davranışsal risk faktörlerinin etkileri bireylerde kan basıncının yükselmesi, kan şekerinin yükselmesi, kan lipidlerinin yükselmesi, fazla kilo ve obezite olarak görülmektedir. Kendini gösteren bu belirtiler kalp krizi, felç, kalp yetmezliği ve diğer komplikasyonların riskinin arttığını göstermektedir. Kardiyovasküler hastalık risklerinin minimuma indirilmesi için diyette tuzun azaltılması, tütün ve alkol kullanımının bırakılması, meyve ve sebze tüketiminin artırılması, düzenli fiziksel aktivite gibi eylemler tavsiye edilmektedir.

Bu önlemlerin yanı sıra hipertansiyon, diyabet ve yüksek kan lipidlerinin ilaç tedavisi kalp hastalıkları riskini azaltmaktadır (URL-3).

C. Kalp Hastalıklarının Belirtileri

Kalp hastalıklarının en önemli belirtileri aritmi, kalp krizi ve kalp yetmezliğidir. Aritmi; göğüste meydana gelen çarpıntı olarak ifade edilmektedir. Kalp krizi belirtileri; göğüs ağrısı veya rahatsızlığı, sırt veya boyun ağrısı, hazımsızlık, mide ekşimesi, mide bulantısı veya kusma, aşırı yorgunluk, üst vücutta rahatsızlık, baş

dönmesi ve nefes darlığıdır. Kalp yetmezliği; nefes darlığı, yorgunluk veya ayaklarda, ayak bileklerinde, bacaklarda, karında veya boyun damarlarında şişmedir (URL-4). Kalp hastalığının erken belirtileri baş dönmesi veya bayılma nöbetleri, yemekten sonra uzun süre devam eden rahatsızlık, nefes darlığı, sebebi açıklanamayan yorgunluk, göğüste ağrı veya sıkışma, göğsün merkezinde uyuşma hissi ve çarpıntı olarak sıralanabilmektedir (Lakshmi vd., 2013).

D. Kalp Hastalıkları Çeşitleri

Kardiyovasküler sistemde, endokardit, romatizmal kalp hastalığı ve iletim sistemi anormalliklerini içeren çok çeşitli problemler ortaya çıkabilmektedir. Kardiyovasküler hastalık, 4 çeşitten oluşmaktadır: koroner kalp hastalığı (KKH) olarak da adlandırılan koroner arter hastalığı (KAH), serebrovasküler hastalık, periferik arter hastalığı (PAH) ve aort aterosklerozudur (AA) (Benjamin vd., 2018).

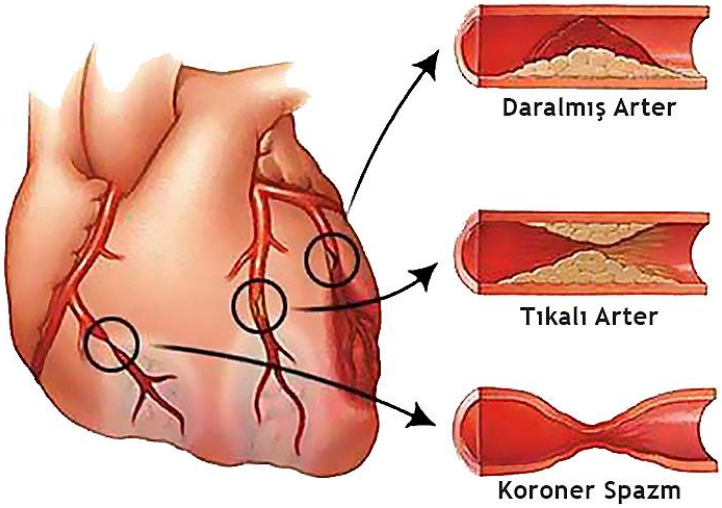
- **Koroner Arter Hastalığı (KAH)**

Türkiye'de ve dünyada ölüm oranları yüksek olan hastalıkların içerisinde bulunan koroner arter hastalıklarının (KAH) risk bulunduran faktörlerini bilinmesi ile toplumun bu hastalıkta hususunda bilinçlendirilmesi koroner arter hastalıklarının engellenmesi için büyük önem taşımaktadır. Koroner arterlerin, yani damar sertliği denilen durumun varlığı sebebi ile tıkalı olması ya da daralmasıdır. Ateroskleroz, atardamarın iç duvarında yağ ile kolesterol birikintilerinin ortaya çıkmasıdır. Damarların duvarın meydana gelen birikintiler plak şeklinde isimlendirilmektedir. Ortaya çıkan bu plaklar arterleri tıkayıp kan dolaşımını önlemekte, arter fonksiyonu ile tonusunda farklılığa sebep olmakta ve kalp kasına gitmekte olan kan akışını kısıtlamaktadır. Kalbe giden kan akışı yeterli seviyede olmadığı takdirde kalp yerine getirmesi gerekli olan hayatsal fonksiyonlarını sürdüremez (Akdemir ve Akyar, 2008).

Koroner arterlerin şekli boş borulara benzemektedir. Bu boruların içerisinde kan serbest bir şekilde akış göstererek dolaşım sağlamaktadır. Normalde koroner arterlerin kas duvarları elastik ve düz bir yapıya sahip olmaktadır. Endotelyum ismi verilmiş olan bu yapı türlü uyarıcılara kimyasal olan sinyaller vererek arterin işlevini düzenlemektedir (Akdemir ve Akyar, 2008).

Türlü etkenlere bağlı şekilde koroner arter hastalığı genç olan kişilerde başlamaktadır. İlk öncesinde kan damarı duvarında yağ çizgileri oluşmaktadır. Yaş

ilerledikçe ortaya çıkan yağ çizgileri yağa dönüştükçe kan damarı duvarlara küçük ölçülerde zarar vermeye başlamaktadır. Kan damarları içerisinde kalsiyum, atık ürünler ile beyaz kan hücreleri gibi maddelerde yer almaktadır ve bu maddeler damar duvarlarına yapışmaya başlamaktadır. Başka maddeler ile yağ birleşince plak ismi verilen maddeyi meydana getirmektedirler (Akdemir ve Akyar, 2008). Şekil 3'de koroner arterdeki oluşan hastalıklara örnek verilmektedir.



Şekil 3 Koroner Arter Hastalığı

Yaş ilerleme gösterdikçe arter içinde değişik büyüklükte ortaya çıkan bu plaklar parçalanabilmekte ve damarların kan akımı yeniden sağlanmış olmaktadır. Birtakım durumlardaysa damarların tamami ile tıkanmasına ve akut koroner sendromlara neden olabilmektedir.

Koroner arter rahatsızlığı bulunuyor ise anjina pektoris ismi verilmiş olan sıkıntı hissi veya göğüs ağrısı olabilmektedir. Kalp kasının çalışmasını sağlamakta olan kan akımı tamami ile kesilmekte kalbin enerji gereksinimi kan akışı miktarından çok daha fazla olduğundan kalp krizi meydana gelebilmektedir. Kalp kriziyle alakalı genel olarak sıkıntı hissi göğüste görülmekte ama bu alanın dışında midenizden dişe ya da çene altına iki koldan kürek kemiğine, el bileğinden parmaklara dek her yerde hissedilebilmektedir. Göğüste ortaya çıkan ağrıya nefes darlığı da eklenebilmekte ayriyeten huzursuzluk, yanma, bulantı, baygınlık, yorgunluk ile kötü bir şey olacak hissi gibi göstergeler de yaşanabilmektedir (Montalescot vd., 2014).

Koroner arter hastalığı için düzeltilemeyen ile düzeltilebilir risk faktörleri bulunmaktadır. Bu riskler Çizelge 1'de verilmektedir.

Çizelge 1 Koroner Arter Hastalığı İçin Düzeltilemeyen ile Düzeltilebilir Risk Faktörleri

Düzeltilemez Risk Faktörleri	**Düzeltilebilir Risk Faktörleri**
Genetik koroner arter hastalığı	Obezite
Erkek olmak	Stres
İlerleyen yaş	Alkol tüketimi
	Yetersiz fiziksel aktivite
	Yüksek kolestrol
	Diyabet
	Yüksek diyet
	Sigara kullanımı

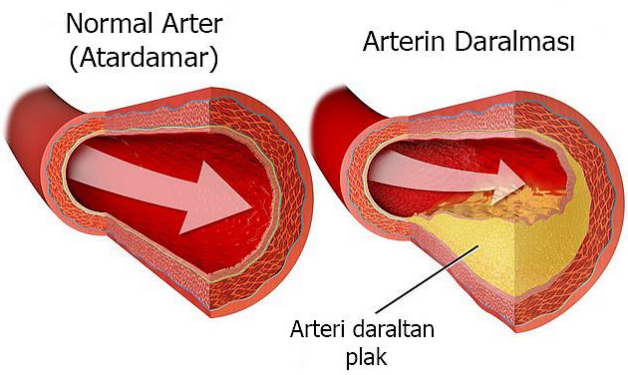
Yukarıda yer alan çizelgede verilen risk faktörleri ne derece fazlaysa hastalık riski de o kadar fazla olmaktadır. Risk faktörleri hususunda kişinin bilgiye sahip olması hayat şeklini değiştirerek kadiyovasküler hastalıklara yakalanma ihtimalinde azalma göstermektedir (Ebrahim vd., 2011).

- **Serebrovasküler Hastalık**

İskemik ile hemorajik olarak serebrovasküler hastalıklar ikiye temel sınıfa ayrılmaktadır.

1. **İskemik serebrovasküler hastalık:** beyni beslemekte olan damlarda tıkanma neticesinde ortaya çıkan nöronal sorunlardır. Çok yaygın görülmekte olan bir inme tipidir. Çok sık görülen iskemik inmelerin sebebi ise aterosklerozdur. Kalsiyum, kolesterol, yağ kan hücreleri ile başka maddelerden meydana gelen plaklar, atardamarlar içerisinde birikim göstererek daralmaya sebep olmaktadırlar. Bu kısımda meydana gelen ani bir kan pıhtılaşması iskemik inmeye sebep olabilmektedir (Öztürk, 2010).
2. **Hemorajik serebrovasküler hastalık:** beyni beslemekte olan damarlarda yırtılma oluşması neticesinde ortaya çıkan nöronal sorunlardır. İskemik inmeye oran ile çok daha az görülmektedir fakat ölümcül olma ihtimali çok daha yükseklik göstermektedir. Beyin içinde zayıflık gösteren bir kan damarının yırtılması neticesinde meydana gelmektedir. Genel olarak bu yırtılmaların nedeni ani tansiyon yükselişleridir. Diğer nedenleri arasındaysa arteriovenöz malformasyonlar (beyin damar yumağı) ile anevrizmalar bulunmaktadır (Öztürk, 2010).

Ölüm sebebi olarak serebrovasküler hastalıklar Dünya'da üçüncü sırada yer alırken, sakatlığa neden olması açısından birinci sırada yer almaktadır. Yaş ile birlikte serebrovasküler hastalıklar artış göstermektedir. 65 yaş üzeri hastalar %75'ini oluşturmaktadır. Bütün inmelerin %60 ile 80'ini iskemik serebrovasküler hastalıklar %10 ile 15'ini hemorajik serebrovasüler hastalık, %3 ile 10'unu subaraknoid kanamalar oluşturmaktadır (Muhsiroğlu, vd., 2017).



Şekil 4 Arterin Normal Hali ve Daralmış Hali

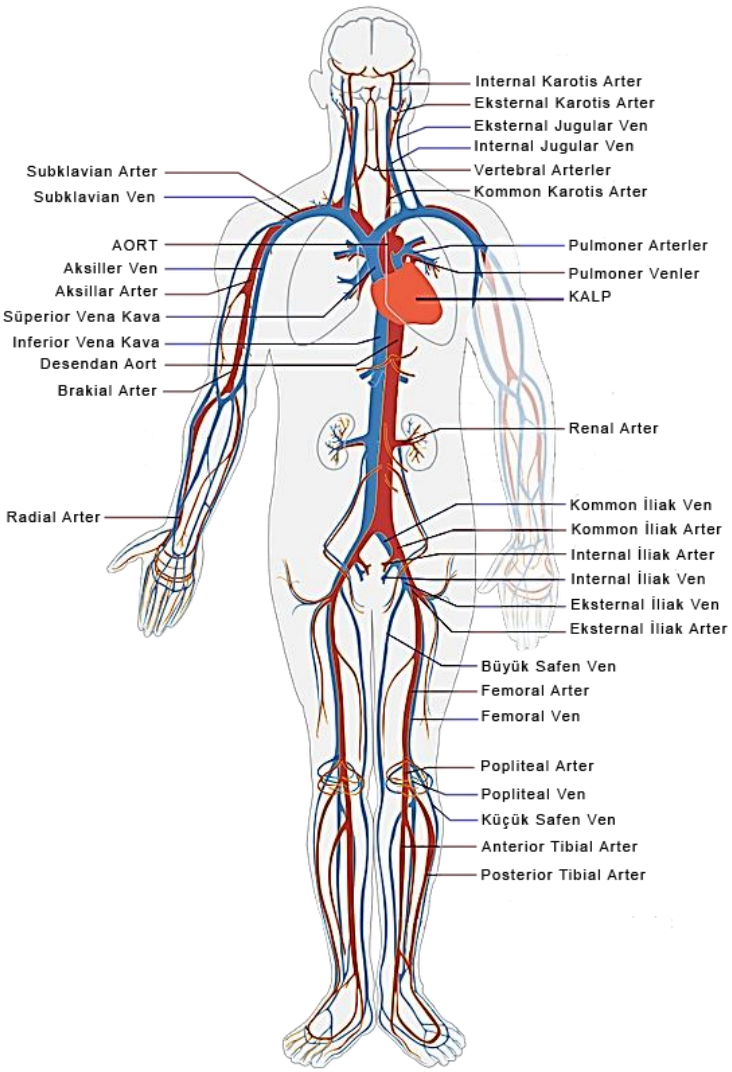
Beynin etki alan yerine bakılarak değişik nörolojik bulgular gelişmektedir. Bu bulgular şu şekilde sıralanabilmektedir:

- Baş dönmesi ve denge kayıpları
- Baş ağrısı
- Hafıza problemleri
- Duyusal hasarlar (Vücudun bir yarısında duyu kaybı)
- Tam ya da kısmi felç
- Görme bozuklukları (bulanık görme, çift görme, yarım görme)
- Konuşma bozuklukları (konuşurken yanlış kelimeler kullanma, konuşulanı anlamada güçlük, konuşamama, sarhoşvari konuşma, peltek konuşma).

Serebrovasküler hastalıkların nedenleri hipertansiyon, kalp hastalıkları, diyabet, hiperlipidemi, sigara, alkol kullanımı, obezite, fiziksel inaktivite, beslenme alışkanlıkları, hiperkoagülabilite, hormon kullanımı, damar hastalıkları, uyuşturuculardır.

- Periferik Arter Hastalığı (PAH)

Vücut içerisinde yer alan aort damarından çıkarak, kol, bacak baş ile organlara temiz olan kanı götürmekte olan atardamar dallarıyla tekrardan bunlardan gelmekte olan kirli kanı kalbe iletmek amacıyla ana toplardamara bağlantılı olan toplardamar dalları periferik damarlar şeklinde isimlendirilmektedir. Bu damarlarda meydana gelen ve kan akışını engelleyecek veya kısıtlayacak hallere periferik damar hastalıkları denilmektedir (Şatiroğlu v., 2011). Aşağıdaki şekilde arter damarlar verilmektedir:



Şekil 5 Vücutta Yer Alan Arterler

Periferik arter hastalığı kol, bacak baş ile organlara kan ileten arterlerde ateroskleroza bağlı şekilde plak oluşmasının neden olduğu hastalık türüdür. Lifli doku, kalsiyum, kolesterol, yağ, plak ile kanda olan başka maddelerden oluşmaktadır.

Periferik arter hastalığı bütün artere etki edebilmektedir fakat en yaygın olduğu şekli bacaklara giden arterlerde görülmektedir. Plak oluşması sebebi ile daralma gösteren atardamardan kan geçişi kısıtlanmakta ve etki altına aldığı alanda ağrılara sebep olmaktadır (Şatiroğlu v., 2011).

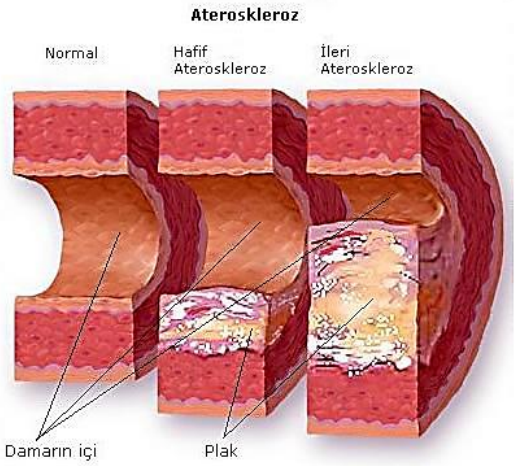
Asıl sebebi aterosklerozdur. Romatizmal hastalıklar, bağışıklık sistemi hastalıkları, inflamasyon, türlü damar yaralanmaları gibi sebeplerden bu evreye hız kazandırabilmektedir. Periferik arter hastalığının ilerlemesi bireyden bireye değişim göstermekle beraber, aile öyküsü, kişinin yaşam tarzı, plağın oluşum yeri, sahip olunmakta olan başka sistematik hastalıklar gibi çok fazla etmene bağlılık göstermektedir.

Başlangıç seviyesinde periferik arter hastalığı herhangi bir belirtiye neden olmamaktadır. Arter içerisinde olan daralma artış gösterdiğinde ve kan akışı kritik biçimde azalma gösterdiğinde, yeterli seviyede kan alamayan dokular ölmeye başlamaktadır.

En yaygın olan bacak arteri hastalığında klodikasyon şeklinde isimlendirilir ve efor ile ortaya bacak ağrıları ile kramplar çıkmaktadır. Merdiven çıkma veya yürüme gibi aktivitelerin ardından baldırlar veya kalçada ağrılı kramplar, bacaklarda ağırlık hissi bacak arterlerinde daralmanın oluştuğuna dair bir göstergedir. Daralma ne derece fazla ise göstergelerin şiddet seviyesi o derece fazla olmaktadır. Damarlarda yaygın olan hastalık halinde, sandalyeye bağlı şekilde bacaklarda morarma, üşüme, soğukluk, bazen uzuv kaybına ve iyileşmeyen yaralara neden olacak doku ölümleri ortaya çıkabilmektedir.

- **Aort Aterosklerozu (AA)**

Damar sertliği şeklinde tanımlanan ateroskleroz, atar damarların iç kısımlarında kolesterol, yağ ile iltihabi atıkların bir araya gelmesi neticesinde ortaya koydukları plaklarla meydana gelen darlık şeklinde açıklanabilmektedir. Bu darlığın neticesinde kan akımı hızı yavaşlar ve organlarda beslenme bozukluğu ortaya çıkmaktadır. Aşağıda yer alan görselde ateroskleroz hastalığının gösterimi bulunmaktadır (Köz, 2016).



Şekil 6 Ateroskleroz Hastalığı

Ateroskleroz hastalığı çok yaygındır ve sistemiktir. Vücutta yalnızca bir alanda lokalize kalmamaktadır. Bu sebep ile damar sertliği bütün atar damarları tutmaktadır. Bu sebep ile ateroskleroza sebep olan ve hastalığının tabanını yapan faktörlere dikkat etmek gereklidir.

Çocukluk yaşlarında başlangıç göstererek yavaşça kendini ortaya koyan bu hastalık gayet sinsi ve karmaşık bir evreyi oluşturmaktadır. Yaşın ilerlemesiyle beraber kendini göstermektedir. Yaygın olarak bu hastalık erkeklerde çok daha fazla gözlemlenmektedir. Genel olarak kadınlarda görülme sıklığı menopozun ardından gelen zamana denk gelmektedir. Bu durumsa östrojen hormonunun azalma göstermesinden kaynaklı olmaktadır (Köz, 2016).

Arterlerin daralmasıyla kan akışının sınırlı hale geldiği ateroskleroz rahatsızlığında duvarın esneklik hali yok olarak sertlik kazanmasına sebep olan bazı faktörler yer almaktadır. Bu faktörler diğer kalp hastalıklarına sebebiyet verenler faktörlerle benzerlik göstermektedir. Faktörler şu şekildedir:

- Yetersiz fizik aktivite
- Sağlıksız beslenme
- Sigara kullanımı
- Şeker hastalığı

- Obezite
- Yüksek tansiyon
- Yüksek kolesterol
- Aile içinde olan kalp – damar hastalıkları

Hastalığın göstergeleri kendine has değildir. Damar sertliğinin kalpte ortaya çıkması durumunda kalp kası zayıflık göstermekte ve kas görevini yeterli derecede yapamamaktadır. Bu durumun neticesinde kalbe giden oksijen seviyesi azalma göstermektedir Ayrıyeten kalpte ritim sorunları da ortaya çıkmaktadır. Ardından kalp krizi ihtimali yükseklik göstermektedir.

E. Kalp Hastalıkları Risk Faktörleri

Bir kişinin hastalık geliştirme olasılığını artıran koşullar veya alışkanlıklar risk faktörleridir. Aynı zamanda risk faktörleri mevcut bir hastalığın kötüleşme olasılığını arttırmaktadır (Lakshmi vd., 2013).

Literatürde kalp hastalıkları risk faktörleri,

- Tütün kullanımı,
- Yetersiz fiziksel aktivite,
- Alkol kullanımı,
- Sağlıklı beslenme,
- Obezite,
- Diabetes mellitus (DM),
- Hipertansiyon,
- Dislipidemi,
- Cinsiyet,
- Yaş,
- Genetik öykü olarak belirlenmiştir (Dülek vd., 2018).

Kalp hastalıkları risk faktörleri değiştirilebilir ve değiştirilemez faktörler olarak ikiye ayrılmaktadır (Türkmen vd., 2010). Değiştirilemez faktörler: yaş, cinsiyet, genetik öykü; değiştirilebilir faktörler: diabetes mellitus ve kötü kan şekeri regülasyonu, sigara kullanımı, hipertansiyon, dislipidemi, obezite veya viseral yağlanma, psikososyal faktörler, sedanter yaşam, fiziksel aktivitenin az olması, meyve ve sebze tüketiminin az olması, düzenli alkol kullanımı olarak sıralanmaktadır (Karakoç vd., 2017). Kalp hastalıklarından korunmak için bu risk faktörlerinin değerlendirilmesi ve tanılanması önleyici açıdan önemlidir (Türkmen vd., 2012).

F. Makine Öğrenmesi

Makine öğrenmesi, bilgisayar biliminin bir alt dalıdır. Makine öğrenmesi teknikleri dünyada hızla yükselen bir konudur. Makine öğrenmesi makinelere verileri daha verimli bir şekilde nasıl kullanacaklarını öğretmek için kullanılmaktadır (Mahesh, 2018). Her geçen gün dünyada veri akışı çoğalmakta ve her gün hızla bu akışa yeni veriler eklenmektedir. CSC'ye göre 2009 - 2020 yılları arasında veri miktarının 44 kat arttığı düşünülmektedir (URL-5). Verilerin işlenmesi ve veriler üzerinde anlamlı tahminlerde bunabilmek için makine öğrenmesi yöntemleri kullanılmaktadır. Makine öğrenmesinin temel amacı geçmiş verilere bakarak gelecek verilere yönelik akılcı tahminlerde bulunmaktır. Oluşturulmuş matematiksel yöntemlerin programlanması ile veriler kolaylıkla analiz edilebilmektedir. Bunun sonucunda dağınık halde ve farklı niteliklerde bulunan verilerden anlamlı sonuçlar üretebilmek mümkün hale gelmiştir.

Birçok sektör, ilgili verileri çıkarmak için makine öğrenimi yöntemlerini kullanmaktadır. Makine öğreniminin amacı verilerden öğrenmektir. Makinelerin kendi kendine öğrenmesini sağlamak için birçok çalışma yapılmış, birçok matematik ve programlama uzmanı çeşitli yaklaşımlar uygulamıştır.

Makine öğrenimi için kullanılan veriler temelde etiketli veri ve etiketsiz veri olarak ikiye ayrılmaktadır. Etiketli veriler, niteliklerin sağlandığı verilerdir. Verilere eklenmiş bir tür etiket veya anlam içermekte olup denetimli öğrenmede kullanılmaktadır. Etiketli nitelik sayısal veya kategorik olabilir. Regresyonda değeri tahmin etmek için sayısal veriler, sınıflandırmada kategorik veriler kullanılmaktadır. Etiketsiz veriler, yalnızca veri noktalarının bulunduğu ve yardımcı olacak etiketlemenin olmadığı verilerdir. Etiketsiz veriler, denetimsiz öğrenmede

kullanılmaktadır (Cunningham vd., 2008: 289). Böylece makine, veri kümesinde bulunan kalıpları veya herhangi bir yapıyı tanımlayabilmektedir.

Etiketli veriler ve etiketsiz veriler sırasıyla denetimli öğrenme ve denetimsiz öğrenme ile kullanılmaktadır. Denetimli öğrenme, bir dizi girdi değişkeni x ile bir çıktı değişkeni y arasında bir öğrenme haritası gerektirmekte ve bilinmeyen veriler için çıktıyı tahmin etmek için bu eşlemeyi uygulamaktadır (Cunningham vd., 2008: 289). Veri kümesini öğrendikten sonra, algoritmalar verileri genelleyerek verilen veri kümesi için varsayımsal H değerini formüle etmektedir. Denetimli öğrenme regresyon ve sınıflandırma olarak iki türe ayrılmaktadır.

Regresyon, iş sözlüğüne göre bağımlı değişkendeki bir değişikliğin ilişkili olduğu ve bir veya daha fazla bağımsız değişkendeki değişikliğe bağlı olduğu iki veya daha fazla değişken arasındaki istatistiksel ilişkiyi belirlemek için kullanılan bir tekniktir (URL-6).

Sınıflandırma, günlük yaşamda çok sık gerçekleşen bir iştir. Nesnelerin her birine sınıflar olarak bilinen karşılıklı olarak kapsamlı ve özel bir dizi kategoriden birine atanacak şekilde bölünmesini içermektedir. Karşılıklı kapsamlı ve dışlayıcı terimi her nesnenin tam olarak bir sınıfa atanması gerektiği anlamına gelmektedir. Nesneler asla birden fazla sınıfa atanmamalı ve bir sınıfa dahil olmalıdır (Bramer, 2013).

Sınıflandırma kategorik değerler üzerinde örüntü kurma, sınıflama ve tahmin etme işlerini yaparken regresyon analizi ise süreklilik veya kesikli durum gösteren değişkenlerin birbirleri arasındaki ilişkiyi belirlemeyi sağlar (Çalış vd., 2014).

Denetimsiz öğrenme, sistemlerin belirli girdi kalıplarını genel girdi kalıpları koleksiyonunun istatistiksel yapısını yansıtacak şekilde temsil etmeyi nasıl öğrenebileceğini inceler. Buna karşılık, denetimli öğrenmede her girdiyle ilişkili açık hedef çıktılar veya çevresel değerlendirmeler bulunmamaktadır. Bunun yerine denetimsiz öğrenen, girdinin yapısının hangi yönlerinin çıktıda yakalanması gerektiğine ilişkin ön önyargılar içermektedir (Wilson, 1999: 858).

G. Biyoinformatik Alanda Makine Öğrenmesi

Biyoinformatik biyolojik verilerin teknoloji kullanılarak işlenmesini sağlayan bir bilim dalıdır. Makine öğrenmesi ve yöntemlerimin gelişimiyle DNA, gen araştırmaları, sinir bilimi, evrimsel biyoloji, ilaç ve kanser araştırmaları gibi konularda biyoinformatik çalışmaları yaygınlaştırılmıştır (Carter vd., 2001). Makine öğrenimi genomik, proteomik mikrodiziler, sistem biyolojisi, evrim ve metin madenciliği alanlarında yaygın olarak kullanılmaktadır (Carter vd., 2001).

Tıp alanında toplanan verilerden hareketle kalp hastalığının teşhis edilmesi üzerine pek çok çalışma bulunmaktadır. Biyolojik verilerin makine öğrenmesi algoritmaları ile incelendiği çalışmaların büyük bır kısmı başarılı sonuçlar vermiştir (Kolay vd. 2016).

Makine öğrenmesinin biyolojik uygulama alanlarına dair kullanılan algoritmaların doğru sonuçlar vermesi ve öğrenme yöntemlerinin doğru seçilmesi önemli bir detaydır. Makine öğrenmesin biyomedikal alanda kardiyovasküler hastalıklar gibi karmaşık hastalıkların teşhisinde kullanılması, erken teşhis gibi önemli konuların göz önünde tutulmasında faydası bulunmaktadır. Bu alandaki çalışmalar, bireysel tedavilerin geliştirilmesinde, kişiye özel ilaç tasarımı uygulamalarında ve kalp hastalıkları gibi ölümcül hastalıkların erken teşhis ve tedavisinde önemli rol oynadığı düşünülmektedir. Ayrıca makine öğrenmesi yöntemlerinin kişiye özel tıp kavramının popüler olduğu bu dönemde kullanılabilecek en yararlı yöntemlerin başında geldiği de düşünülmektedir (Baldi ve Brunak, 2001).

IV. YÖNTEM VE TEKNİKLER

Bu tez çalışmasında, veri seti üzerinde veri işleme teknikleri kullanılarak Python 3.10.5. sürümü ile optimizasyon yapılıp Yapay Sinir Ağları, Destek Vektör Makineleri (SVM), k-NN (En Yakın Komşu), Rastgele Orman (Random Forest) ve Naive Bayes ana makine öğrenmesi modelleri ile WEKA 3.6.12. ortamında sınıflandırma uygulanmıştır. Sonuçlar hedef değişken için ayrı ayrı değerlendirilerek en başarılı sonuçları üreten sınıflandırma algoritmaları başarıları karşılaştırılmıştır.

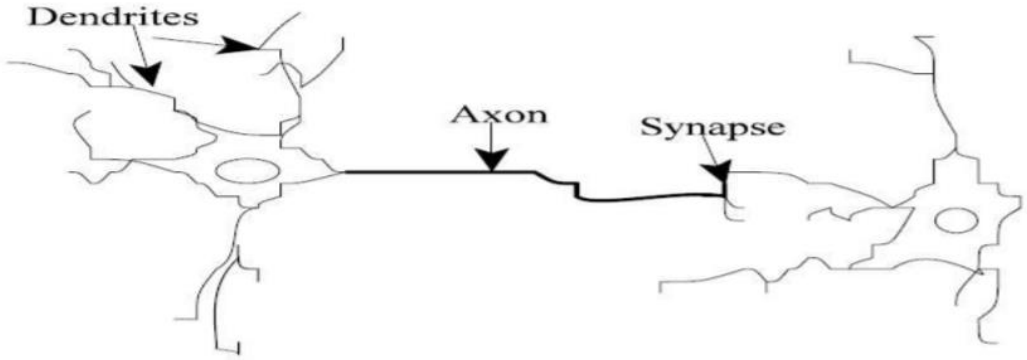
A. Makine Öğrenmesi Algoritmaları

Makine öğrenmesi algoritmaları, insanların karmaşık veri kümelerini keşfetmesine, analiz etmesine ve bunlarda anlam bulmasına yardımcı olan kod parçacıklardır. Her algoritma, bir makinenin belirli bir hedefi gerçekleştirmek için izleyebileceği sınırlı ve belirli adım adım ilerleyen yönerge kümesidir. Makine öğrenmesi modelinin hedefi, insanların tahmin yapmak veya bilgileri kategorilere ayırmak için kullanabileceği desenler oluşturmak veya keşfetmektir (URL-8). Makine Öğrenimi, veri sorunlarını çözmek için farklı algoritmalara dayanır. Veri bilimcileri, bir sorunu çözmek için en iyi olan, her duruma uyan tek bir algoritma türü olmadığını belirtmekten hoşlanırlar. Kullanılan algoritma türü, çözmek istediğiniz problemin türüne, değişkenlerin sayısına, ona en uygun modelin türüne bağlıdır (Mahesh, 2018). Makine öğrenmesi algoritmaları denetimli öğrenme, denetimsiz öğrenme ve pekiştirmeye dayalı öğrenme tekniklerini kullanmaktadır.

1. Yapay Sinir Ağları (YSA)

İnsan beyni, çok karmaşık sorunları çözebilen, oldukça karmaşık bir makinedir. YSA'yı anlamak için, beynin iç kısımlarının nasıl çalıştığına dair temel bilgilere sahip olmamız gerekmektedir. Beyin, merkezi sinir sisteminin bir parçasıdır ve çok büyük bir sinir ağından (NN) oluşur. (Nissen, 2003)

NN, birbirine bağlı nöronlardan oluşan bir ağdır. Nöronun merkezine çekirdek denir. Çekirdek, dendritler ve akson aracılığıyla diğer çekirdeklere bağlanır. Bu bağlantıya sinaptik bağlantı denir. Nöron, diğer nöronların dendritlerinde alınan sinaptik bağlantıları aracılığıyla elektrik darbeleri ateşleyebilir. Şekil 7, basitleştirilmiş bir nöronun nasıl göründüğünü gösterir.

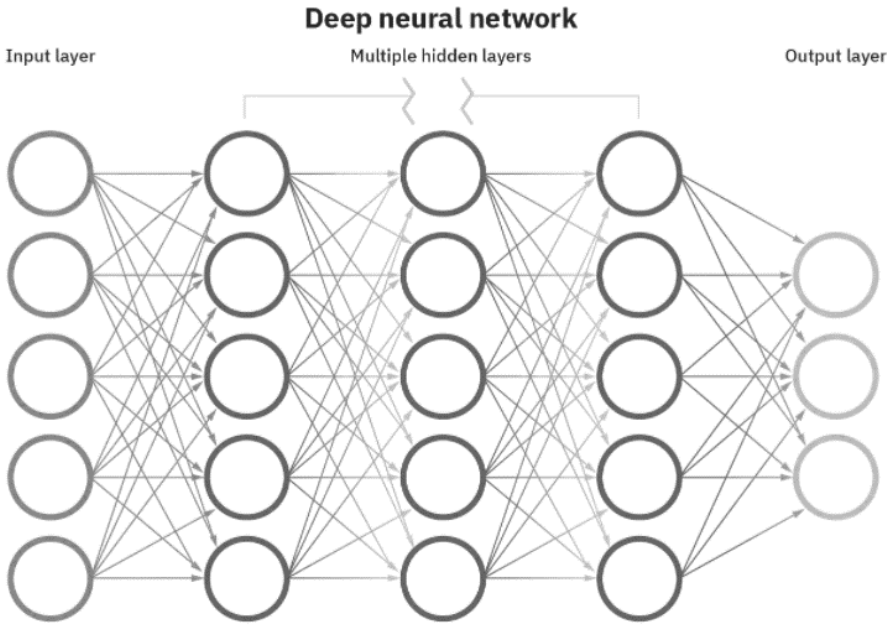


Şekil 7 Basitleştirilmiş Nöron (Zakaria vd., 2014)

Bir nöron, dendritleri aracılığıyla yeterli elektrik darbesi aldığında, aksonu aracılığıyla bir darbeyi aktive eder ve ateşler, bu daha sonra diğer nöronlar tarafından alınır. Bu şekilde bilgi NN aracılığıyla yayılabilir. Bir nöronun ömrü boyunca sinaptik bağlantılar değişir ve bir nöronu (eşik) etkinleştirmek için gereken gelen darbelerin miktarı da değişir. Bu davranış, NN'nin öğrenmesini sağlar. İnsan beyni, yaklaşık 1015 bağlantıyla yüksek düzeyde bağlantılı olan yaklaşık 1011 nörondan oluşur (Tettamanzi ve Tomassini, 2001). Bu nöronlar iç ve dış kaynaklara bir etki olarak paralel olarak aktive olurlar. Beyin, sinir sisteminin geri kalanıyla bağlantılıdır, bu da onun beş duyu aracılığıyla bilgi almasını ve aynı zamanda kasları kontrol etmesini sağlar.

Günümüz teknolojisinde basitleştirilmiş yapay nöronlar ve yapay sinir ağları yapmak mümkün hale gelmiştir. Bu YSA'lar birçok farklı şekilde yapılabilir ve beyni birçok farklı şekilde taklit etmeye çalışabilir. YSA'lar akıllı değildir, ancak kalıpları tanımak ve karmaşık problemler için basit kurallar oluşturmak için iyidirler. Ayrıca mükemmel eğitim yeteneklerine sahiptirler, bu nedenle yapay zeka araştırmalarında sıklıkla kullanılırlar. YSA'lar bir dizi eğitim verisinden genelleme yapmakta iyidir. (Nissen, 2003)

Yapay sinir ağları (YSA), bir girdi katmanı, bir veya daha fazla gizli katman ve bir çıktı katmanı içeren bir düğüm katmanından oluşur. Her düğüm veya yapay nöron diğerine bağlanır ve ilişkili bir ağırlık ve eşiğe sahiptir. Herhangi bir düğümün çıktısı belirtilen eşik değerinin üzerindeyse, o düğüm etkinleştirilir ve ağın bir sonraki katmanına veri gönderilir. Aksi takdirde, ağın bir sonraki katmanına hiçbir veri iletilmez (IBM, 2020).



Şekil 8 Yapay Sinir Ağı (IBM, 2020)

Nörona gönderilen her girdi önce ağırlıklandırılmalı, -1 ile 1 arasında bir sayı çarpılmalıdır. Bir algılayıcı oluşturmak genellikle rastgele ağırlıklar atayarak başlar. Her girdi alınır ve ağırlığı ile çarpılır. Algılayıcının çıktısı, bu toplamın bir aktivasyon fonksiyonundan geçirilmesiyle üretilir. Aralarından seçim yapabileceğiniz birçok aktivasyon fonksiyonu bulunmaktadır (Acharya, 2017).

Sinir ağları, zaman içinde doğruluklarını öğrenmek ve geliştirmek için eğitim verilerine güvenir. Bununla birlikte, bu öğrenme algoritmaları doğruluk için ince ayar yapıldığında, bilgisayar bilimi ve yapay zekada güçlü araçlardır ve verileri yüksek bir hızda sınıflandırmamıza ve kümelememize olanak tanır. En iyi bilinen sinir ağlarından biri Google'ın arama algoritmasıdır.

Yapay Sinir Ağı, biyolojik sinir ağlarından sonra modellenmiştir ve bilgisayarların insan destekli öğrenmeye benzer şekillerde öğrenmesine izin vermeye çalışır. Algılayıcı; sinir ağlarında çalışan temel birimdir. Bir algılayıcı, bir veya daha

fazla girdi, bir işlemci ve tek bir çıktıdan oluşur. Bir algılayıcı ileri besleme modelini takip eder, yani girdiler bir nörona gönderilir, işlenir ve çıktı ile sonuçlanır.

Yapay Sinir Ağı olarak adlandırılan bir sinir ağının en basit tanımı, ilk nörobilgisayarlardan biri olan Dr. Robert Hecht-Nielsen'in mucidi tarafından yapılmıştır. Maureen'in "Neural Network Primer: Part I" adlı kitabında Yapay Zeka Uzmanı Caudill sinir ağını "...bir dizi basit, birbirine yüksek düzeyde bağlı işlem öğelerinden oluşan, bilgileri harici girdilere dinamik durum yanıtlarıyla işleyen bir bilgi işlem sistemi" olarak tanımlamaktadır.

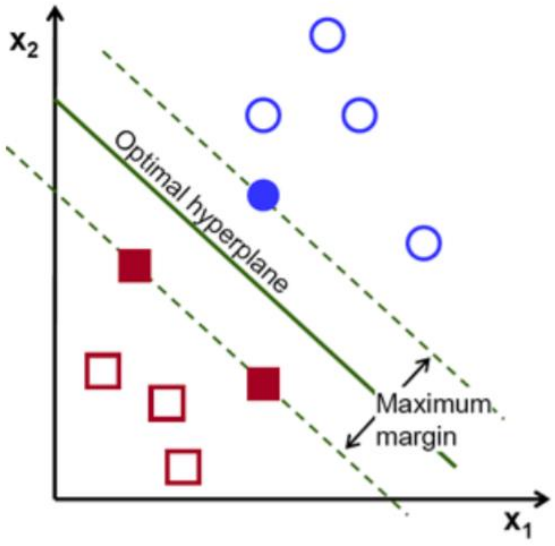
YSA'lar, insan serebral korteksinin nöronal yapısından sonra gevşek bir şekilde modellenen ancak çok daha küçük ölçeklerde modellenen işleme cihazlarıdır. Büyük bir YSA yüzlerce veya binlerce işlemci birimine sahip olabilirken, insan beyninde genel etkileşimlerinin ve ortaya çıkan davranışların büyüklüğünde buna karşılık gelen milyarlarca nöron bulunur.

Sinir ağları ile ilgili matematik önemsiz bir konu olmasa da, bir kullanıcı, yapı ve işlevleri hakkında en azından operasyonel bir anlayış elde edebilir (URL-9).

2. Destek Vektör Makineleri (DVM)

Destek Vektör Makinesi (SVM), 1990'lı yıllarda önerilen ve çoğunlukla örüntü tanıma için kullanılan en iyi makine öğrenme algoritmalarından biridir. Birçok disiplinde veri ayrımında güçlü bir araçtır. SVM, denetimli bir makine öğrenimi türüdür. Her biri birçok kategoriden birine ait olarak işaretlenmiş bir dizi eğitim örneği verildiğinde, bir SVM eğitim algoritmasının yeni örneğin kategorisini tahmin eden bir model oluşturduğu algoritmadır. DVM, istatistiksel öğrenmede amaç olan sorunu genelleştirme konusunda daha fazla yeteneğe sahiptir.

Destek Vektör Makinesi (SVM), bir hiper düzlem kullanarak çeşitli veri sınıflarını ayıran bir sınıflandırıcıdır. SVM, eğitim verileriyle modellenir ve test verilerinde hiper düzlemin çıktısını verir. SVM modeli, farklı veri sınıflarının geniş ölçüde farklılaştırılabileceği veri matrisindeki alanı bulmaya çalışır ve bir hiper düzlem çizer.



Şekil 9 Destek Vektör Makinesi (Ponraj vd., 2020)

Şekil 9'da Kırmızı ve Mavi, etiketlenmiş eğitim veri noktalarının sınıflarıdır. Bunları lineer olarak sınıflandırmak için bir hiper düzlem çizilebilir. Bir hiper düzlem çizmenin birden fazla yolu vardır. Sınıflar arasındaki marjı maksimize eden optimal bir hiper düzlem seçilir. Hiper düzlemin her zaman doğrusal olması gerekmez. SVM'deki bir hiper düzlem, çekirdek hilesi olarak bilinen tekniği kullanarak doğrusal olmayan bir sınıflandırıcı olarak da çalışabilir.

3. En Yakın Komşu (k-NN)

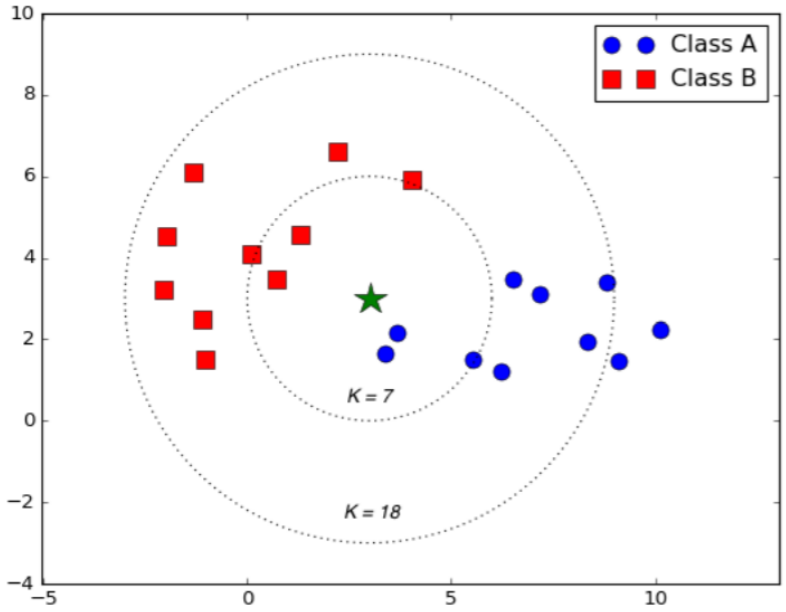
En Yakın Komşu algoritmaları, tüm makine öğrenimi algoritmalarının en basitleri arasındadır. Buradaki fikir, eğitim setini ezberlemek ve ardından eğitim setindeki en yakın komşularının etiketlerine dayanarak herhangi bir yeni örneğin etiketini tahmin etmektir. Böyle bir yöntemin arkasındaki mantık, etki alanı noktalarını tanımlamak için kullanılan özelliklerin, yakın noktaların aynı etikete sahip olmalarını sağlayacak şekilde etiketlemeleriyle ilgili olduğu varsayımına dayanmaktadır. Ayrıca, bazı durumlarda, eğitim seti çok büyük olduğunda bile, en yakın komşuyu bulmak son derece hızlı bir şekilde yapılabilir (Shai, 2014: 258)

K-NN temel olarak, bir mesafe fonksiyonuna göre verilerin sınıflandırılmasına ve en yakın değerler ile eşleştirilmesine dayanır. Yeni veriler eğitim veri setindeki komşulara olan yakınlıklarına göre test edilir ve mesafe fonksiyonunda belirlenen k değerine göre uygun sınıfa alınır. K = 1 ise, test verisi en yakın komşusunun sınıfına atanır.

K-NN algoritması için kullanılan çeşitli mesafe fonksiyonları bulunmaktadır. Sürekli değişkenler için kullanılan 3 ayrı mesafe fonksiyonu tanımlanmıştır (Lu ve Zu, 2014)

Öklid Mesafe Ölçümü: $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$ (Denklem 1)

k-en yakın komşudaki k, yeni örneğe en yakın veri noktalarının sayısıdır. Örneğin, k =1 ise algoritma en yakın örneği seçecek veya k = 4 ise algoritma en yakın dört komşu örneği seçecek ve bunları buna göre sınıflandıracaktır. Fikir, Şekil 10 ile daha iyi gösterilebilir.



Şekil 10 En Yakın Komşu (k) (URL-11)

Şekil 10'a bakıldığında, Yeşil yıldız sınıflandırılacak veri noktasıdır, mavi daireler A sınıfı veri noktalarıdır ve kırmızı kareler B sınıfı dikdörtgenlerdir. Yeşil yıldız ile diğer tüm kırmızı ve mavi noktalar arasındaki Öklid mesafesi ölçülür. Yıldız, mesafenin en az olduğu veri noktalarına göre sınıflandırılacaktır. k =7 ise, yıldızdan itibaren yedi noktanın tümü arasındaki mesafe ölçülür ve yıldız, bu durumda mavi veri noktası ile en az mesafeli veri noktalarına sınıflandırılır.

4. Naive Bayes

Naive Bayes yöntemi, Thomas Bayes'in (1702-1761) çalışmasına dayanmaktadır (Panda vd., 2007). Bayes sınıflandırmasında, verilen verilerin belirli bir sınıfa ait olduğu hipotezi vardır. Daha sonra hipotezin doğru olma olasılığını

hesaplanır. Bu, belirli sorun türleri için en pratik yaklaşımlardan biridir. Yaklaşım, tüm verilerin yalnızca bir kez taranmasını gerektirir. Ayrıca, eğer bir aşamada ek eğitim verileri varsa, o zaman her eğitim örneği, bir hipotezin doğru olma olasılığını aşamalı olarak artırabilir ya da azaltabilir. Bu nedenle, belirsizlik içeren bir alanı modellemek için bir Bayes ağı kullanılır (Jenson, 2001).

NB algoritmasında olasılıkların hesaplanması için alternatif birçok formül üretilmiştir (URL 12).

$$P(A \mid B) = \frac{P(B \mid A)}{P(A)\,P(B)} \quad \text{(Denklem 2)}$$

P (A | B) = B'nin gerçekleşmesi durumunda A'nın gerçekleşme olasılığı

P (A) = A'nın gerçekleşme olasılığı

P (B | A) = A'nın gerçekleşmesi durumunda B'nin gerçekleşme olasılığı

P (B) = B'nin gerçekleşme olasılığı

Üç tür Naive Bayes vardır. Gauss Naïve Bayes, Multinomial Naïve Bayes ve Bernoulli Naïve Bayes. Sınıflandırma problemlerinde Gaussian Naive Bayes, çok terimli dağıtılmış verilerde Multinomial Naive Bayes ve çok değişkenli Bernoulli dağılımına sahip verilerde Bernoulli Naive Bayes kullanılmaktadır.

5. Lojistik Regresyon

Lojistik Regresyon temelinde bağımlı değişkenin ikili değişken olması durumunda kullanılan bir istatistiksel yöntemdir. Klasik doğrusal regresyon modelinde bağımlı değişken ikili bir değişken olduğunda bazı sorunlar ortaya çıkar. Bu sorunlardan biri değişen varyanstır. Değişen varyans modelin hata teriminin varyansının yanlı tahmin edilmesine ve dolayısıyla yapılan tüm istatistiksel hipotez testlerinde güvenilir olmayan sonuçlar elde edilmesine sebep olur. Her ne kadar bu problemin üstesinden gelinebilse de nitel bağımlı değişkenli modellerde klasik doğrusal regresyon modeli ile tahmin yapmak sebebiyle ortaya çıkan ve üstesinden gelinemeyen problemler de bulunmaktadır. Bu problemlerden biri ve hatta en önemlisi hesaplanan koşullu olasılıkların sıfır bir aralığının dışına çıkmasıdır. Bir diğer problem ise En Küçük Kareler yönteminin varsayımlarından biri olan hataların normal dağılıma sahip olması varsayımının sağlanmamasıdır. Tüm bu problemlerden dolayı ortaya çıkan logit model kümülatif olasılık dağılım fonksiyonundan türetilmiş matematiksel

bir fonksiyondur. Bu fonksiyon aracılığıyla olasılıklar hesaplanarak 0.50'nin üzerindeki olasılıklara 1, diğerlerine sıfır verilerek sınıflandırma yapılır.

6. Karar Ağaçları ve Rassal Ormanlar

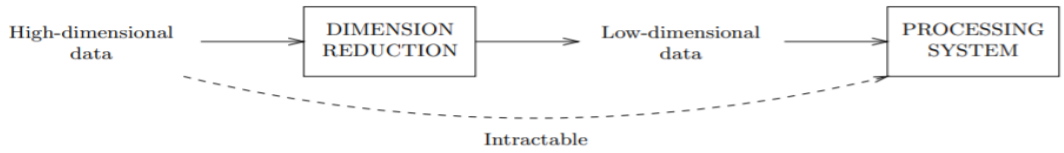
Karar ağaçları algoritması denetimli öğrenmede en sık kullanılan algoritmalardan biridir. Sınıflandırma ve regresyon problemleri için kullanılmaktadır. Kategorik ve sürekli bağımlı değişkenler için kullanılabilir. Algoritma, karar ve yaprak düğümleri olan bir ağaç olarak nitelendirilebilir. Herhangi bir yaprak, sınıflandırma veya kararı temsil eder. Bu şekilde kök düğümde bulunan karardan yapraklara doğru sınıflandırmalar yapılır. Birden fazla karar ağacı algoritmasının birleşmesiyle ise Rassal Orman algoritması oluşur. Bu algoritma sınıflandırma ve regresyon problemlerinde sıklıkla kullanılmaktadır. Özellikle biyolojik hesaplamalarda kullanılmaktadır.

B. Boyutsal Küçültme Teknikleri

Boyutsal küçültme, bilgi kaybı olmadan girdi rastgele değişkenlerinin sayısını azaltma işlemidir (Acharya, 2017). Birçok durumda, ilk önce verilerin boyutunu yönetilebilir bir boyuta indirgemek, orijinal bilgileri mümkün olduğu kadar fazla tutmak ve ardından küçültülmüş boyutlu verileri sisteme beslemek gerekli olabilir. Yüksek boyutlu bir uzayda temsil edilen verileri manipüle etmede daha fazla verimlilik elde etmek için, genellikle boyutu önemli ölçüde azaltmak gerekir.

Daha fazla sayıda girdi değişkeni veya boyutu ve büyük veri örnekleri, veri kümesinin karmaşıklığını artırır. Belleği ve hesaplama zamanını azaltmak için verilerin boyutsallığı azaltılır. Boyut azaltma, aynı zamanda, yinelenen değişkenler veya çok düşük önem düzeyine sahip değişkenler gibi gereksiz girdi değişkenlerinin ortadan kaldırılmasına da yardımcı olur (Alapydin, 2010: 110)

Şekil 11'de tüm sistemde bir ön işleme aşaması olarak boyut küçültmeyi gösteren bu durumu özetlemektedir (Carreira-Perpiñán, 1997).



Şekil 11 Veri Ön İşleme Aşaması

Daha fazla sayıda girdi değişkeni veya boyutu ve büyük veri örnekleri, veri kümesinin karmaşıklığını artırır. Belleği ve hesaplama zamanını azaltmak için verilerin boyutsallığı azaltılır. Boyut azaltma, aynı zamanda, yinelenen değişkenler veya çok düşük önem düzeyine sahip değişkenler gibi gereksiz girdi değişkenlerinin ortadan kaldırılmasına da yardımcı olur (Alapydin, 2010: 109-110) .

Boyut küçültme sorunu, veriler aslında tolere edilenden daha yüksek bir boyutta olduğunda ortaya çıkar. Özellik Seçimi ve Özellik Çıkarma olmak üzere iki tür boyutluluk azaltma tekniği vardır:

1. Özellik Seçimi

Özellik seçiminde, en fazla bilgiyi veren d boyutundan k boyutu seçilir ve (d-k) boyutları atılır. Diğer bir deyişle, özellik seçimine alt küme seçimi de denir. En iyi alt küme, doğruluğa en fazla katkıda bulunan en az sayıda boyutu içerir. En iyi alt küme, uygun hata fonksiyonu ile bulunur (Alapydin, 2010: 110).

Eğitim verileri belirli bir alt küme oluşturma sürecinden geçirilir. Ortaya çıkan alt küme, performansın beklenen kriterleri karşılaması durumunda performansını test etmek için algoritmadan geçirilir, ardından son alt küme olarak seçilecektir. Aksi takdirde, ortaya çıkan alt küme, daha fazla ince ayar için alt küme oluşturma sürecinden geçirilecektir.

Denetimli tümevarımsal öğrenme açısından, özellik seçimi, üç yaklaşımdan birini kullanarak bir dizi aday özellik verir (Molina vd., 2002)

- Bir değerlendirme ölçüsünü optimize eden özelliklerin alt kümesinin belirtilen boyutu
- Değerlendirme ölçütlerinde belirli bir kısıtlamayı karşılayan alt kümenin daha küçük boyutu
- Genel olarak, büyüklük ve değerlendirme ölçütü arasında en iyi bağlılığa sahip alt küme

Bu nedenle, öznitelik seçme algoritmalarının öznitelik seçimi için doğru kullanımı, ya genelleme kapasitesi, öğrenme hızı açısından ya da indüklenen modelin karmaşıklığını azaltma açısından tümevarımsal öğrenmeyi geliştirir.

Eğitim verileri, örneğin sıralı geriye doğru seçim gibi belirli bir alt küme oluşturma sürecinden geçirilir. Ortaya çıkan alt küme, performansın beklenen kriterleri karşılaması durumunda performansını test etmek için şimdi algoritmadan geçirilir, ardından son alt küme olarak seçilir. Aksi takdirde, ortaya çıkan alt küme, daha fazla ince ayar için alt küme oluşturma sürecinden geçirilecektir.

Öznitelik seçimi sürecinde, verilerdeki alakasız ve fazlalık özellikler veya gürültü, mikrodizi veri analizi gibi sınıf kavramı ile ilgili ve önemli olmadıkları için birçok durumda engellenebilir (Dash vd., 1997). Örnek sayısı özelliklerden çok daha az olduğunda, arama alanı seyrek doldurulacağından makine öğrenimi özellikle zorlaşır. Bu nedenle model, gürültü ile ilgili veriler arasında doğru bir ayrım yapamayacaktır (Provost, 2000). Özellik seçimi için iki ana yaklaşım vardır. Birincisi Bireysel Değerlendirme, ikincisi ise Alt Küme Değerlendirmesidir. Özelliklerin sıralanması Bireysel Değerlendirme olarak bilinir (Gyon vd., 2003). Bireysel Değerlendirmede, bireysel bir özelliğin ağırlığı, ilgililik derecesine göre atanır. Alt Küme Değerlendirmesinde, aday özellik alt kümeleri, arama stratejisi kullanılarak oluşturulur.

Özellik seçimi için genel prosedür, dört temel adıma sahiptir.

- Altküme Üretimi
- Alt Kümenin Değerlendirilmesi
- Durdurma Kriterleri
- Sonuç Doğrulama

Öznitelik seçimi için Sıralı İleri Seçim ve Sıralı Geri Seçim olmak üzere iki farklı yaklaşım bulunmaktadır.

- **Sıralı İleri Seçim**

Sıralı İleri Seçim, hiçbir tahmin edici içermeyen bir modelle başlar ve daha sonra, tahmin edicilerin tümü modele girene kadar, tahmin ediciler birer birer modele eklenir. Her adımda uyum için en büyük ek iyileştirmeyi sağlayan değişken modele eklenir.

Xi, i = 1,……,d değişkenleriyle P ile bir küme gösterelim. E(P), test örneğinde oluşan hatadır. Sıralı İleri Seçim, değişkensiz boş küme ile başlar P = {φ}. Her adımda,

boş kümeye tek bir değişken eklenir ve bir model eğitilir ve test kümesinde hata E(P∪ Xi) hesaplanır. Hata kriterleri, örneğin en küçük kare hatası ve yanlış sınıflandırma hatası gibi gereksinime göre belirlenir. Tüm hatalardan en az hataya neden olan Xj girdi değişkeni seçilir ve boş P kümesine eklenir. Model, kalan değişken sayısı ile yeniden eğitilir ve E(P∪ Xi) ise süreç P'ye değişkenler eklemeye devam eder.

- **Sıralı Geri Seçim**

Sıralı Geri Seçim, en iyi alt küme çözümü için verimli bir alternatiftir, ancak Sıralı İleri Seçimden farklı olarak tam özellik kümesiyle başlar. En az önemli özellikleri yinelemeli olarak birer birer kaldırır. Ardışık Geri Seçim, tam değişken seti P = {1,2,3,…..,d} ile başlar. Her adımda model tam bir değişken seti ile eğitilir ve test setinde hata hesaplanır ve en yüksek Xj hatasına sahip değişken P setinden çıkarılır. Model yeni bir değişken seti P ile tekrar eğitilir ve işlem, E(P-Xj) E(P)'den küçükse, değişkenleri P'den çıkarmaya devam eder.

2. Özellik Çıkarma

Öznitelik çıkarma tekniğinde, veri kümesindeki öznitelikler veya bağımsız değişkenler, yeni öznitelik uzayı olarak bilinen yeni bağımsız değişkenlere dönüştürülür. Özellik Çıkarma, mevcut özelliklerden yeni özellikler oluşturup orijinal özellikleri atarak bir veri kümesindeki özelliklerin sayısını azaltmayı amaçlar. Yeni oluşturulan özellik alanı verileri daha iyi açıklar. Yalnızca önemli veriler seçilir. Öznitelik çıkarma yöntemlerinin amacı, daha fazla analiz yapmayı mümkün kılmak için verilerin gereğinden fazla uydurulmasından kaçınmaktır.

Öznitelik çıkarma, ham verinin orijinal veri setindeki bilgiler korunurken işlenebilecek sayısal özelliklere dönüştürülmesi sürecini ifade eder. Makine öğrenimini doğrudan ham verilere uygulamaktan daha iyi sonuçlar verir.

Özellik çıkarma işlemi, önemli veya ilgili bilgileri kaybetmeden işleme için gereken kaynak sayısını azaltmanız gerektiğinde yararlıdır. Özellik çıkarma, belirli bir analiz için gereksiz veri miktarını da azaltabilir. Ayrıca, verilerin azaltılması ve makinenin değişken özellikler oluşturma çabaları, makine öğrenmesi sürecinde öğrenme ve genelleme adımlarının hızını kolaylaştırır.

Özellik çıkarma, manuel veya otomatik olarak gerçekleştirilebilir: (URL-7)

Manuel özellik çıkarma, belirli bir problemle ilgili özellikleri tanımlamayı ve bu özellikleri çıkarmanın bir yöntem uygulamayı gerektirir. Birçok durumda, arka plan veya etki alanı hakkında iyi bir anlayışa sahip olmak, hangi özelliklerin yararlı olabileceği konusunda bilinçli kararlar verilmesine yardımcı olabilir.

Otomatik özellik çıkarma, insan müdahalesine gerek kalmadan sinyallerden veya görüntülerden özellikleri otomatik olarak çıkarmak için özel algoritmalar veya derin ağlar kullanır. Bu teknik, ham verilerden makine öğrenimi algoritmaları geliştirmeye hızlı bir şekilde geçmek istediğinizde çok faydalı olabilir.

Derin öğrenmenin yükselişiyle birlikte, çoğunlukla görüntü verileri için özellik çıkarmanın yerini büyük ölçüde derin ağların ilk katmanları almıştır. Sinyal ve zaman serisi uygulamaları için, öznitelik çıkarma, etkili tahmine dayalı modeller oluşturmadan önce önemli uzmanlık gerektiren ilk zorluk olmaya devam etmektedir.

Öznitelik çıkarma tekniğinde, veri kümesindeki öznitelikler veya bağımsız değişkenler, yeni öznitelik uzayı olarak bilinen yeni bağımsız değişkenlere dönüştürülür.

Yeni oluşturulan özellik alanı, verileri en çok açıklar ve yalnızca önemli veriler seçilir. X1…….Xn'nin n özelliği iken özellik çıkarımından sonra, (n > m) olan m öznitelik vardır ve bu öznitelik çıkarımı bazı eşleme işlevi F ile yapılır.

Xn bağımsız özellikler veya boyutlar kümesi, Yn bağımsız özellikler kümesine indirgenir. Öznitelik çıkarma işleminde temel bileşen analizi gibi bir teknik kullanılır. Xn'den yalnızca yedekli olmayan ve önemli öznitelikleri alacak ve yeni öznitelik uzayı Yn'e dönüşecektir. Öznitelik çıkarımı ile, öznitelik çıkarımından sonra elde edilen Yn öznitelikleri Xn ile aynı olmadığı için, Xn'nin doğrudan bir alt kümesi olmadığı için yorumlama yeteneği kaybolur.

C. Veri Ön İşleme

Modern dünyada çeşitli kaynaklardan toplanan çok miktarda veri bulunmaktadır. Ancak verilerin çoğu kontrollü ortamdan alınmadığı için eksik değerler, gürültüler ve hatalar barındırabilmektedir. Verilerin en uygun şekilde modellenmesi doğruluğu şüpheli olan veriler için yanlış çıktılar oluşturmaktadır. Bu nedenle veri ön işleme, veri analizi ve makine öğreniminin etkin ve doğru kullanımı önemli bir iştir. Veri ön işleme ham verileri anlaşılır bir formata dönüştüren bir dizi

prosedürdür. Aykırı değerlerin tespit edilmesini, aykırı değerlerle ne yapılacağına karar verilmesini, uygun eksik değerlerin bulunmasını ve doldurulmasını ve verilerdeki tutarsızlığın aranmasını sağlamaktadır. Var olan verinin makine öğrenimi algoritmalarına uygulanmadan önce normalleştirilmesi, standartlaştırılması ve azaltılması gerekir. Veri setinin özellikleri farklı ölçüm birimlerine sahip olduğunda verilerin normalleştirilmesi yapılır. Standardizasyon, verileri ölçeklendirmek için kullanılır. Veri kümesinde gereksiz ya da sonucu etkilemeyen özellikler olduğunda, özellik seçimi olarak adlandırılan verilerin azaltılması işlemi yapılmaktadır.

1. Veri Seti

Davranışsal Risk Faktörü Gözetim Sistemi (BRFSS), CDC tarafından 1984 yılından beri yürütülen yıllık olarak toplanan sağlıkla ilgili bir telefon anketidir. Anket, her yıl 400.000'den fazla Amerikalıdan sağlıkla ilgili risk davranışları, kronik sağlık koşulları ve önleyici hizmetlerin kullanımı hakkında yanıtlar toplamaktadır. Bu tez çalışmasında 2015 yılına ait veri seti kullanılmıştır. Oriijinal veri seti 441.455 kişiden gelen yanıtları içermektedir. 330 özelliğe sahiptir. Bu özellikler doğrudan sorulan sorular ya da hesaplanan değişkenlerdir.

BRFSS verileri makine öğrenimi algoritmaları için kullanılabilir bir formatta temizlenmiştir. 330 sütundan oluşan veri kümesinde kalp hastalığını ve diğer kronik sağlık koşullarını etkileyen faktörlere ilişkin kalp hastalığı araştırmalarına dayanarak, bu analize yalnızca belirli özellikler dahil edilmiştir.

Alanda yapılan araştırmalar, aşağıdakileri kalp hastalığı ve diyabet gibi diğer kronik hastalıklar için önemli risk faktörleri olarak tanımlamıştır.

- Kan basıncı (yüksek)
- Kolesterol (yüksek)
- Sigara içmek
- Şeker hastalığı
- Obezite
- Yaş
- Seks
- Yarış

- Diyet
- Egzersiz yapmak
- Alkol tüketimi
- Bmı
- Hane geliri
- Medeni hal
- Uyumak
- Son kontrolden bu yana geçen süre
- Eğitim
- Sağlık sigortası
- Akıl sağlığı

Diyabet ve Kalp Hastalığı sonuçları, diyabetikler için birincil ölüm nedeninin kalp hastalığı komplikasyonları olmasıyla güçlü bir şekilde ilişkilidir (Xie vd., 2019).

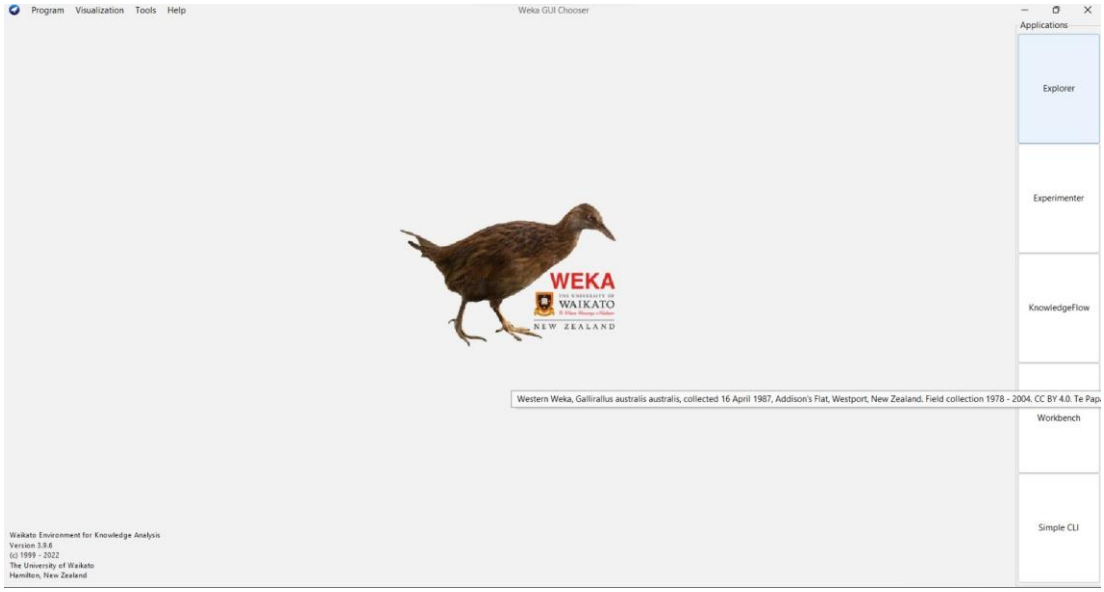
Temizlenen veri seti BRFSS 2015'ten 253.680 anket yanıtı içerir. 229.787 katılımcının kalp hastalığı yoktur, 23.893 kişinin ise kalp hastalığı vardır.

D. WEKA

WEKA açık kaynak kodlu, GNU ile GPL lisansına sahip, Java tabanlı bir makine öğrenmesi aracıdır. Ücretsiz olarak kullanılabilen bir yazılımdır. Yeni Zelanda'da bulunan Waikato Üniversitesi tarafından geliştirilmiştir. WEKA, grafik kullanıcı arayüzleri ile birlikte veri analizi ve tahmine dayalı modelleme için bir dizi görselleştirme araçları ve algoritmaları içermektedir. WEKA ile veri ön işleme, kümeleme, sınıflandırma, ilişkilendirme, görselleştirme ve özellik seçimi gibi temel veri madenciliği işlemleri yapılabilmektedir (Aher vd, 2011). Bu işlemlerin yanı sıra aynı veriye birden fazla algoritma uygulanabilmektedir. WEKA öğrenme şemalarının istatiksel olarak değerlendirilmesine olanak sağlamaktadır.

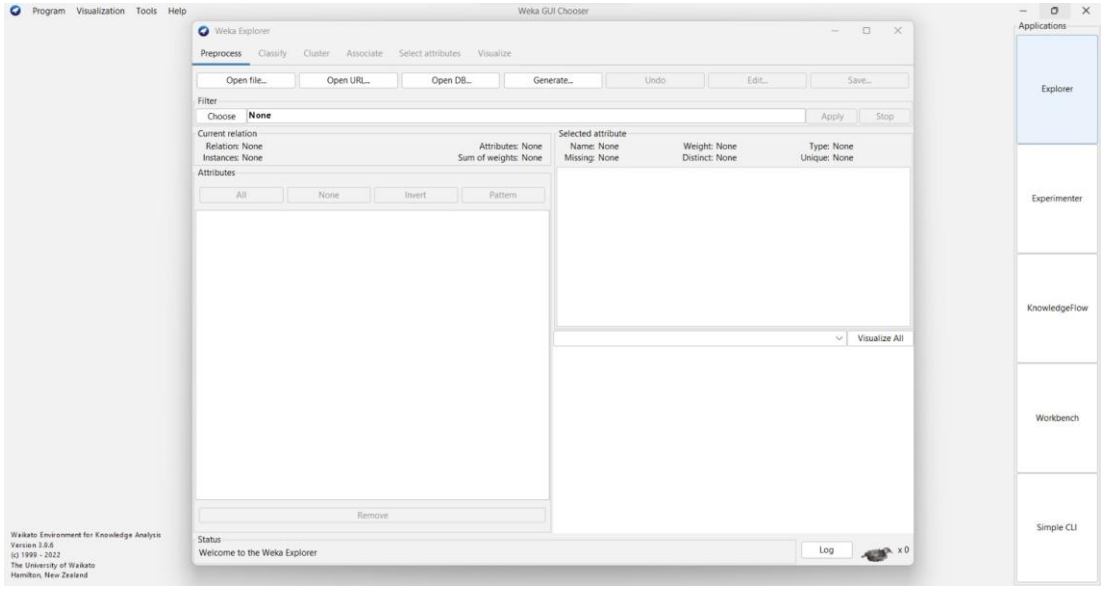
WEKA kullanıcılara yapılacak işlemlerin niteliğine göre Explorer, Experimenter, Knowladge Flow, Work Bench, Simple CLI olmak üzere farklı arayüzlere sahiptir (Witten vd., 2016).

Weka paket programının giriş ekranı Şekil 12'de verilmiştir.



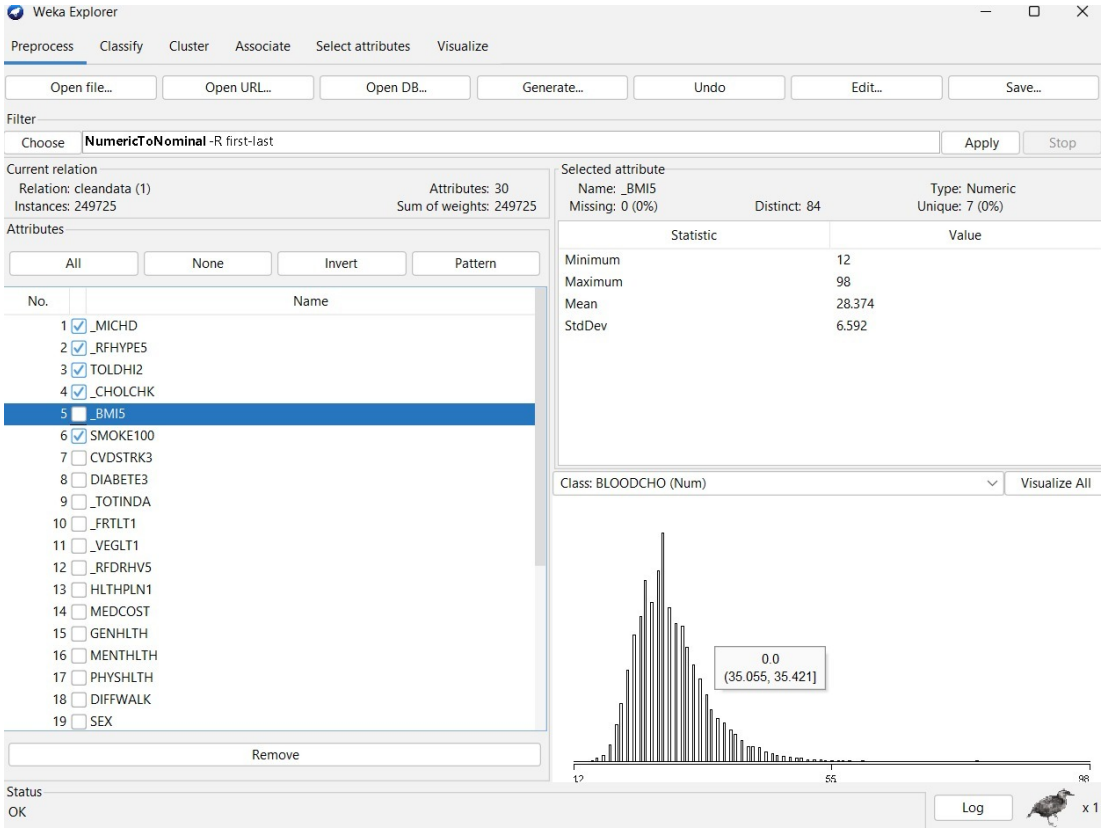
Şekil 12 Weka Programının Giriş Ekranı

Weka programının veri seçimi ekranı Şekil 13'de gösterilmiştir.



Şekil 13 Weka Programında Veri Seçim Ekranı

Weka programının veri ön işleme ekranı Şekil 14'de verilmiştir.

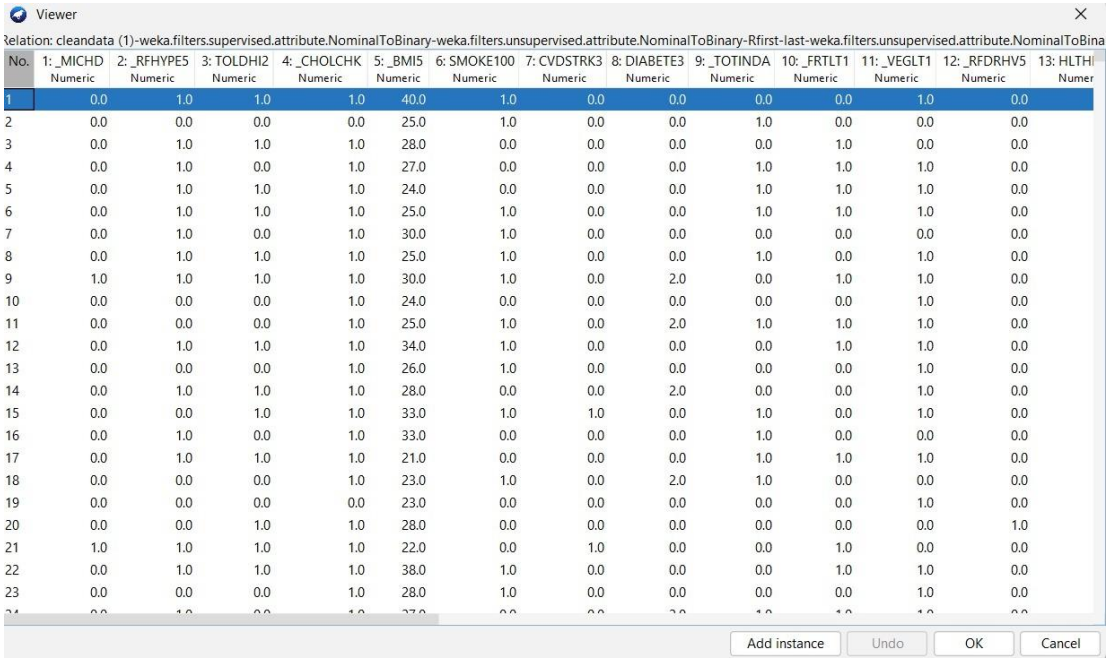


Şekil 14 Weka Programında Veri Ön İşleme Ekranı

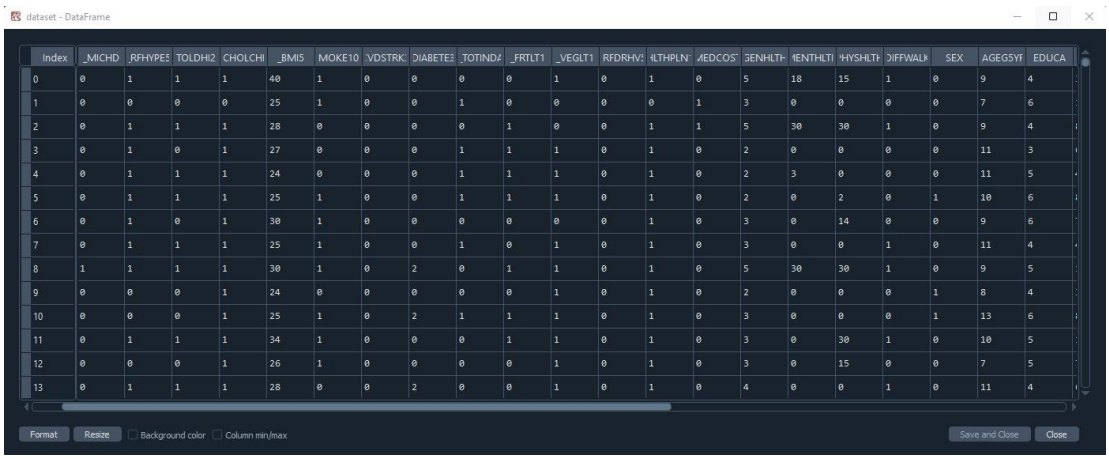
1. Veri Ön İşleme

- **Veri Ön İşleme Süreci**

Verinin ilk hali hem Python hem de Weka paket programında Şekil 15 ve Şekil 16'daki gibidir.

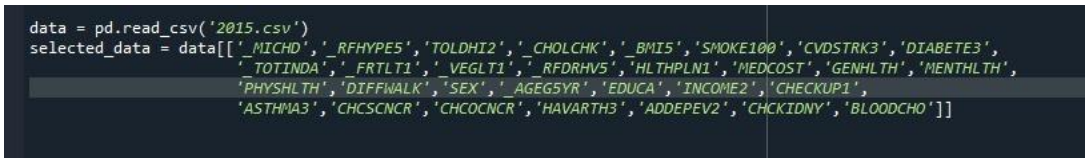


Şekil 15 Weka Programında Verilerin İlk Hali



Şekil 16 Python Programında Verilerin İlk Hali

Öncelikle kalp rahatsızlığına sahip olma ile alakalı değişkenler seçilmiştir, 330 değişken bulunan veri seti 30 değişkene düşürülmüştür.



```
data = pd.read_csv('2015.csv')
selected_data = data[['_MICHD','_RFHYPE5','TOLDHI2','_CHOLCHK','_BMI5','SMOKE100','CVDSTRK3','DIABETE3',
                      '_TOTINDA','_FRTLT1','_VEGLT1','_RFDRHV5','HLTHPLN1','MEDCOST','GENHLTH','MENTHLTH',
                     'PHYSHLTH','DIFFWALK','SEX','_AGEG5YR','EDUCA','INCOME2','CHECKUP1',
                     'ASTHMA3','CHCSCNCR','CHCOCNCR','HAVARTH3','ADDEPEV2','CHCKIDNY','BLOODCHO']]
```

Şekil 17 Kullanılan Veri Seti

Bu işlem sonrasında kayıp gözlemler veri setinden çıkartılmıştır, değişkenler nominal veya ikili değişken olduğu için eksik gözlemleri ortalama ile doldurmak mümkün değildir.

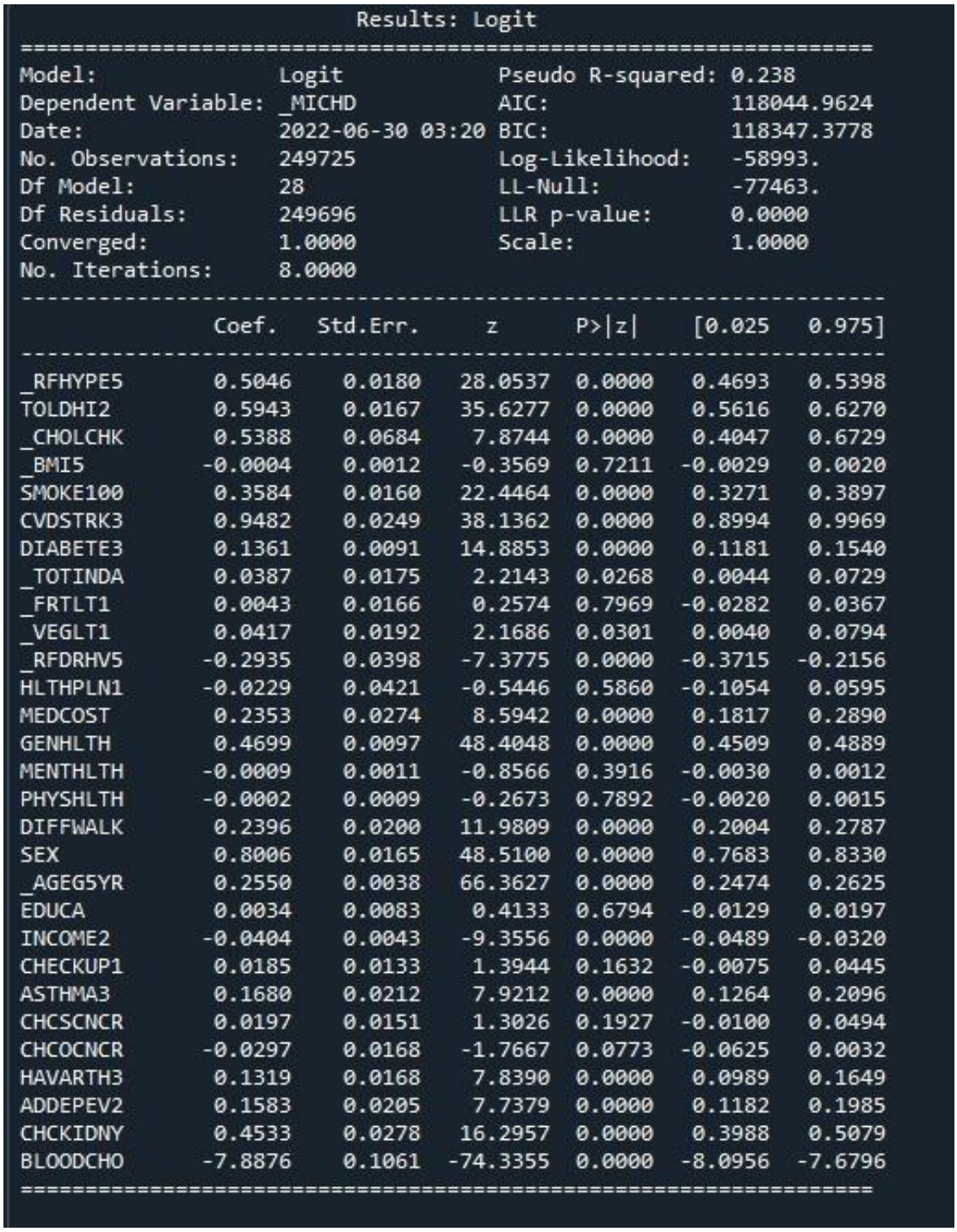
Ardından tüm değişkenler için 7 ve 9 değerini alan gözlemler ile 77 ve 99 değerini alan gözlemler veri setinden çıkartılmıştır çünkü bu gözlemler "Cevap vermedi" ya da "Kararsız" cevaplarını temsil etmektedir.

Değişkenlerin ölçeklendirilme süreci için yapılan işlemler şöyledir:

- CHECKUP1 adlı değişken üzerinde 1 ve 2 değerini alan gözlemler 0 olarak, 3 ve 4 değerini alan gözlemler 1 olarak değiştirilmiştir.
- BLOODCHO adlı değişken için 2 değerini alan gözlemler 0 olarak değiştirilmiştir.
- ASTHMA3 adlı değişken için 2 değerini alan gözlemler 0 olarak değiştirilmiştir.
- CHCSCNCR adlı değişken için 2 değerini alan gözlemler 0 olarak değiştirilmiştir.
- CHCOCNCR adlı değişken için 2 değerini alan gözlemler 0 olarak değiştirilmiştir.
- HAVARTH3 adlı değişken için 2 değerini alan gözlemler 0 olarak değiştirilmiştir.
- ADDEPEV2 adlı değişken için 2 değerini alan gözlemler 0 olarak değiştirilmiştir.
- CHCKIDNY adlı değişken için 2 değerini alan gözlemler 0 olarak değiştirilmiştir.
- _MICHD adlı değişken için 2 değerini alan gözlemler 0 olarak değiştirilmiştir.
- _RFEYEE5 adlı değişken için 1 değerini alan gözlemler 0 ve 2 değerini alan gözlemler 1 olarak değiştirilmiştir.
- TOLDPI2 adlı değişken için 2 değerini alan gözlemler 0 olarak değiştirilmiştir.
- _CHOLCHK adlı değişken için 2 ve 3 değerini alan gözlemler 0 olarak değiştirilmiştir.
- _BMI5 değişkeni yüzdelik hale getirilmek için 100'e bölünmüştür.
- SMOKE100 adlı değişken için 2 değerini alan gözlemler 0 olarak değiştirilmiştir.

- CVDSTRK3 adlı değişken için 2 değerini alan gözlemler 0 olarak değiştirilmiştir.
- DIABETE3 adlı değişken için 2 ve 3 değerini alan gözlemler 0, 1 değerini alan gözlemler 2 ve 4 değerini alan gözlemler 1 olarak değiştirilmiştir.
- _TODINDA adlı değişken için 2 değerini alan gözlemler 0 olarak değiştirilmiştir.
- _FRTLT1 adlı değişken için 2 değerini alan gözlemler 0 olarak değiştirilmiştir.
- _VEGLT1 adlı değişken için 2 değerini alan gözlemler 0 olarak değiştirilmiştir.
- _RFDRHV5 adlı değişken için 1 değerini alan gözlemler 0 ve 2 değerini alan gözlemler 1 olarak değiştirilmiştir.
- HLTHPLN1 adlı değişken için 2 değerini alan gözlemler 0 olarak değiştirilmiştir.
- MEDCOST adlı değişken için 2 değerini alan gözlemler 0 olarak değiştirilmiştir.
- MENTHLTH adlı değişken 88 değerini alan gözlemler 0 olarak değiştirilmiştir.
- PSYCHLTH adlı değişken 88 değerini alan gözlemler 0 olarak değiştirilmiştir.
- DIFFWALK adlı değişken için 2 değerini alan gözlemler 0 olarak değiştirilmiştir.
- SEX adlı değişken için 2 değerini alan gözlemler 0 olarak değiştirilmiştir.

Geriye doğru eleme yöntemi kullanılarak yapılan özellik (öznitelik) seçimi yapılmıştır. Seçim yapılmadan önceki model Şekil 18'de verilmiştir.



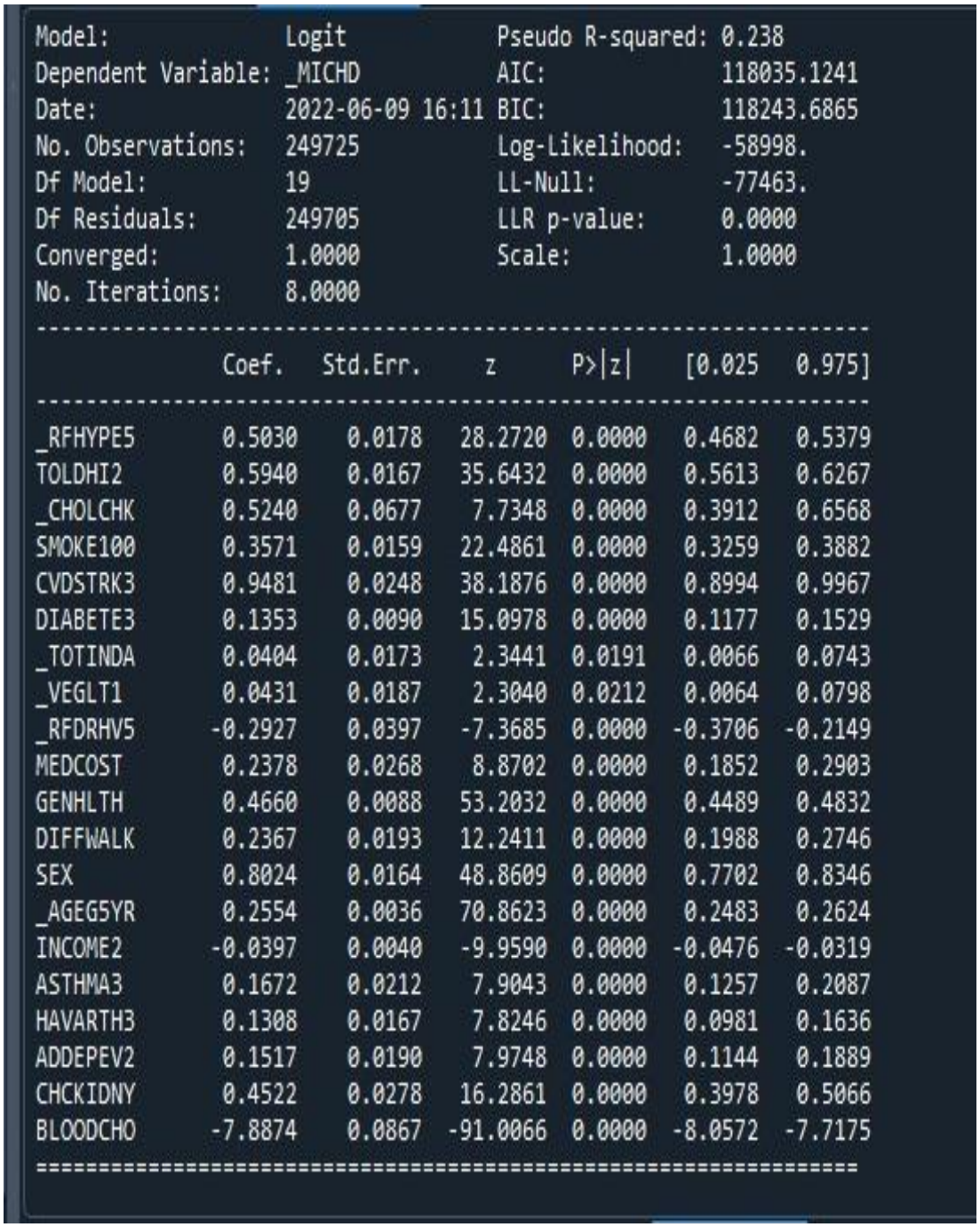
Results: Logit

Model:	Logit	Pseudo R-squared:	0.238
Dependent Variable:	_MICHD	AIC:	118044.9624
Date:	2022-06-30 03:20	BIC:	118347.3778
No. Observations:	249725	Log-Likelihood:	-58993.
Df Model:	28	LL-Null:	-77463.
Df Residuals:	249696	LLR p-value:	0.0000
Converged:	1.0000	Scale:	1.0000
No. Iterations:	8.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
_RFHYPE5	0.5046	0.0180	28.0537	0.0000	0.4693	0.5398
TOLDHI2	0.5943	0.0167	35.6277	0.0000	0.5616	0.6270
_CHOLCHK	0.5388	0.0684	7.8744	0.0000	0.4047	0.6729
_BMI5	-0.0004	0.0012	-0.3569	0.7211	-0.0029	0.0020
SMOKE100	0.3584	0.0160	22.4464	0.0000	0.3271	0.3897
CVDSTRK3	0.9482	0.0249	38.1362	0.0000	0.8994	0.9969
DIABETE3	0.1361	0.0091	14.8853	0.0000	0.1181	0.1540
_TOTINDA	0.0387	0.0175	2.2143	0.0268	0.0044	0.0729
_FRTLT1	0.0043	0.0166	0.2574	0.7969	-0.0282	0.0367
_VEGLT1	0.0417	0.0192	2.1686	0.0301	0.0040	0.0794
_RFDRHV5	-0.2935	0.0398	-7.3775	0.0000	-0.3715	-0.2156
HLTHPLN1	-0.0229	0.0421	-0.5446	0.5860	-0.1054	0.0595
MEDCOST	0.2353	0.0274	8.5942	0.0000	0.1817	0.2890
GENHLTH	0.4699	0.0097	48.4048	0.0000	0.4509	0.4889
MENTHLTH	-0.0009	0.0011	-0.8566	0.3916	-0.0030	0.0012
PHYSHLTH	-0.0002	0.0009	-0.2673	0.7892	-0.0020	0.0015
DIFFWALK	0.2396	0.0200	11.9809	0.0000	0.2004	0.2787
SEX	0.8006	0.0165	48.5100	0.0000	0.7683	0.8330
_AGEG5YR	0.2550	0.0038	66.3627	0.0000	0.2474	0.2625
EDUCA	0.0034	0.0083	0.4133	0.6794	-0.0129	0.0197
INCOME2	-0.0404	0.0043	-9.3556	0.0000	-0.0489	-0.0320
CHECKUP1	0.0185	0.0133	1.3944	0.1632	-0.0075	0.0445
ASTHMA3	0.1680	0.0212	7.9212	0.0000	0.1264	0.2096
CHCSCNCR	0.0197	0.0151	1.3026	0.1927	-0.0100	0.0494
CHCOCNCR	-0.0297	0.0168	-1.7667	0.0773	-0.0625	0.0032
HAVARTH3	0.1319	0.0168	7.8390	0.0000	0.0989	0.1649
ADDEPEV2	0.1583	0.0205	7.7379	0.0000	0.1182	0.1985
CHCKIDNY	0.4533	0.0278	16.2957	0.0000	0.3988	0.5079
BLOODCHO	-7.8876	0.1061	-74.3355	0.0000	-8.0956	-7.6796

Şekil 18 Seçim Yapılmadan Önceki Model

Bu modelden %5 önem düzeyinde istatistiksel olarak anlamsız olan FRTL1, PYSHLTH, BMI5, EDUCA, HLTHPLN1, CHCOCNCR, CHECKUP1 ve CHCSCNCR değişkenleri modelden dışlanarak nihai model elde edilmiştir.

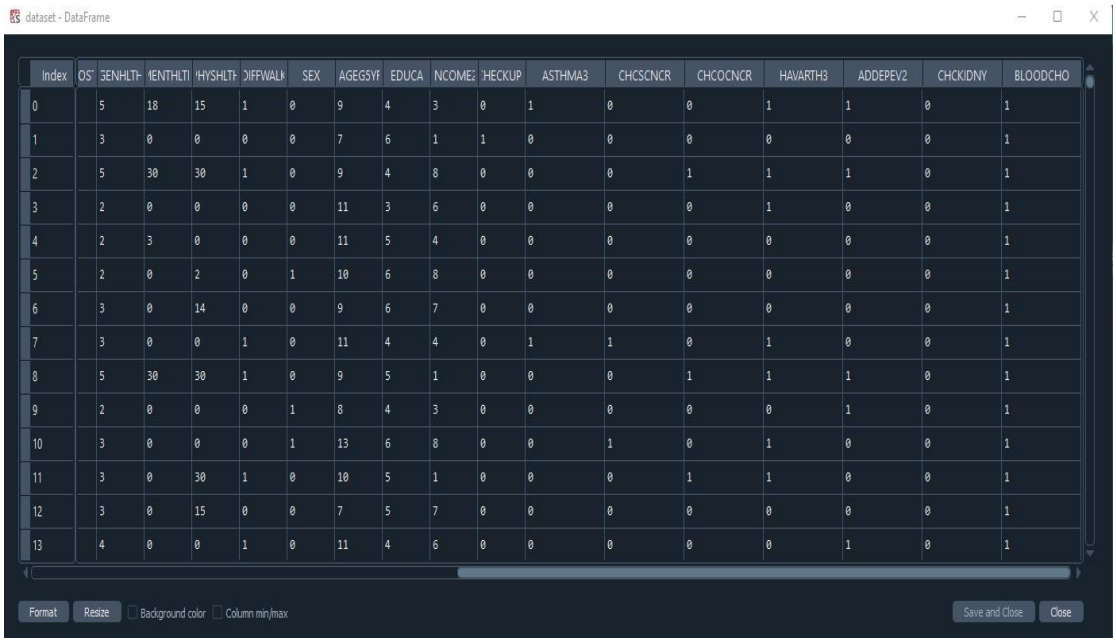


Model:	Logit	Pseudo R-squared:	0.238
Dependent Variable:	_MICHD	AIC:	118035.1241
Date:	2022-06-09 16:11	BIC:	118243.6865
No. Observations:	249725	Log-Likelihood:	-58998.
Df Model:	19	LL-Null:	-77463.
Df Residuals:	249705	LLR p-value:	0.0000
Converged:	1.0000	Scale:	1.0000
No. Iterations:	8.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
_RFHYPE5	0.5030	0.0178	28.2720	0.0000	0.4682	0.5379
TOLDHI2	0.5940	0.0167	35.6432	0.0000	0.5613	0.6267
_CHOLCHK	0.5240	0.0677	7.7348	0.0000	0.3912	0.6568
SMOKE100	0.3571	0.0159	22.4861	0.0000	0.3259	0.3882
CVDSTRK3	0.9481	0.0248	38.1876	0.0000	0.8994	0.9967
DIABETE3	0.1353	0.0090	15.0978	0.0000	0.1177	0.1529
_TOTINDA	0.0404	0.0173	2.3441	0.0191	0.0066	0.0743
_VEGLT1	0.0431	0.0187	2.3040	0.0212	0.0064	0.0798
_RFDRHV5	-0.2927	0.0397	-7.3685	0.0000	-0.3706	-0.2149
MEDCOST	0.2378	0.0268	8.8702	0.0000	0.1852	0.2903
GENHLTH	0.4660	0.0088	53.2032	0.0000	0.4489	0.4832
DIFFWALK	0.2367	0.0193	12.2411	0.0000	0.1988	0.2746
SEX	0.8024	0.0164	48.8609	0.0000	0.7702	0.8346
_AGEG5YR	0.2554	0.0036	70.8623	0.0000	0.2483	0.2624
INCOME2	-0.0397	0.0040	-9.9590	0.0000	-0.0476	-0.0319
ASTHMA3	0.1672	0.0212	7.9043	0.0000	0.1257	0.2087
HAVARTH3	0.1308	0.0167	7.8246	0.0000	0.0981	0.1636
ADDEPEV2	0.1517	0.0190	7.9748	0.0000	0.1144	0.1889
CHCKIDNY	0.4522	0.0278	16.2861	0.0000	0.3978	0.5066
BLOODCHO	-7.8874	0.0867	-91.0066	0.0000	-8.0572	-7.7175

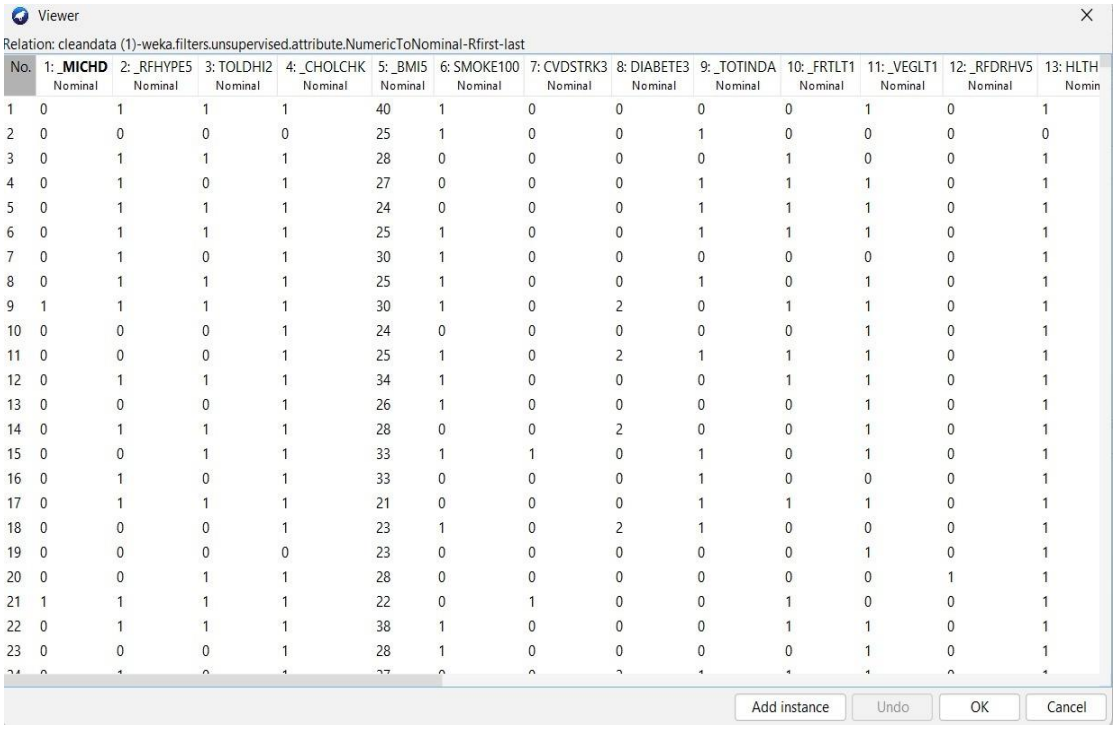
Şekil 19 Nihai Model

Veri setinin son hali Şekil 20 ve 21'de verilmiştir.



Index	OS	GENHLTH	MENTHLTH	PHYSHLTH	DIFFWALK	SEX	AGEG5YR	EDUCA	INCOME2	CHECKUP	ASTHMA3	CHCSCNCR	CHCOCNCR	HAVARTH3	ADDEPEV2	CHCKIDNY	BLOODCHO
0		5	18	15	1	0	9	4	3	0	1	0	0	1	1	0	1
1		3	0	0	0	0	7	6	1	1	0	0	0	0	0	0	1
2		5	30	30	1	0	9	4	8	0	0	0	1	1	1	0	1
3		2	0	0	0	0	11	3	6	0	0	0	0	1	0	0	1
4		2	3	0	0	0	11	5	4	0	0	0	0	0	0	0	1
5		2	0	2	0	1	10	6	8	0	0	0	0	0	0	0	1
6		3	0	14	0	0	9	6	7	0	0	0	0	0	0	0	1
7		3	0	0	1	0	11	4	4	0	1	1	0	1	0	0	1
8		5	30	30	1	0	9	5	1	0	0	0	1	1	1	0	1
9		2	0	0	0	1	8	4	3	0	0	0	0	0	1	0	1
10		3	0	0	0	1	13	6	8	0	0	1	0	1	0	0	1
11		3	0	30	1	0	10	5	1	0	0	0	1	1	0	0	1
12		3	0	15	0	0	7	5	7	0	0	0	0	0	0	0	1
13		4	0	0	1	0	11	4	6	0	0	0	0	0	1	0	1

Şekil 20 Veri Setinin Son Hali-1 (Python)

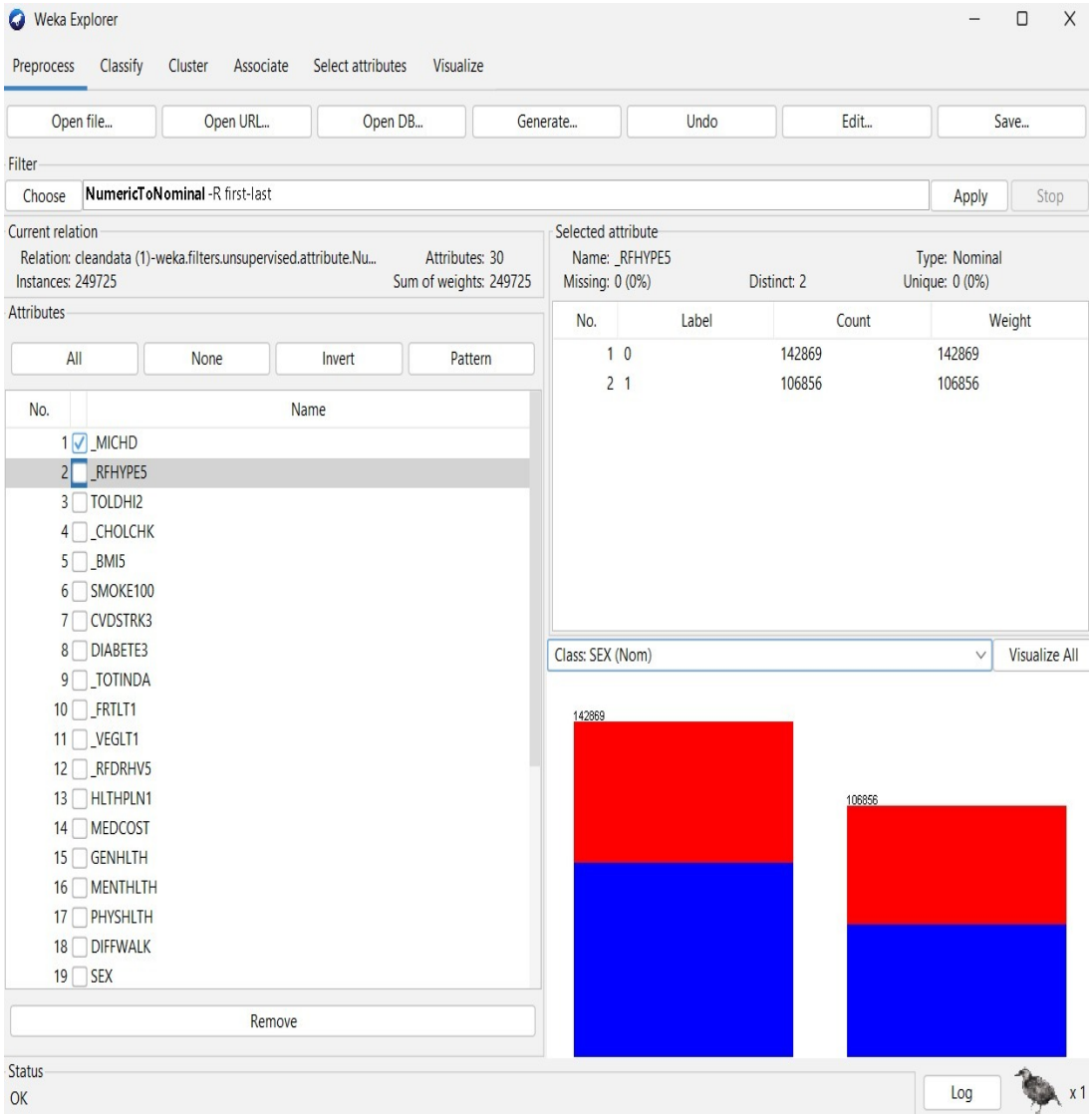


Relation: cleandata (1)-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last

No.	1: _MICHD Nominal	2: _RFHYPE5 Nominal	3: TOLDHI2 Nominal	4: _CHOLCHK Nominal	5: _BMI5 Nominal	6: SMOKE100 Nominal	7: CVDSTRK3 Nominal	8: DIABETE3 Nominal	9: _TOTINDA Nominal	10: _FRTLT1 Nominal	11: _VEGLT1 Nominal	12: _RFDRHV5 Nominal	13: HLTH Nomin
1	0	1	1	1	40	1	0	0	0	0	1	0	1
2	0	0	0	0	25	1	0	0	1	0	0	0	0
3	0	1	1	1	28	0	0	0	0	1	0	0	1
4	0	1	0	1	27	0	0	0	1	1	1	0	1
5	0	1	1	1	24	0	0	0	1	1	1	0	1
6	0	1	1	1	25	1	0	0	1	1	1	0	1
7	0	1	0	1	30	1	0	0	0	0	0	0	1
8	0	1	1	1	25	1	0	0	1	0	1	0	1
9	1	1	1	1	30	1	0	2	0	1	1	0	1
10	0	0	0	1	24	0	0	0	0	0	1	0	1
11	0	0	0	1	25	1	0	2	1	1	1	0	1
12	0	1	1	1	34	1	0	0	0	1	1	0	1
13	0	0	0	1	26	1	0	0	0	0	1	0	1
14	0	1	1	1	28	0	0	2	0	0	1	0	1
15	0	0	1	1	33	1	1	0	1	0	1	0	1
16	0	1	0	1	33	0	0	0	1	0	0	0	1
17	0	1	1	1	21	0	0	0	1	1	1	0	1
18	0	0	0	1	23	1	0	2	1	0	0	0	1
19	0	0	0	0	23	0	0	0	0	0	1	0	1
20	0	0	1	1	28	0	0	0	0	0	0	1	1
21	1	1	1	1	22	0	1	0	0	1	0	0	1
22	0	1	1	1	38	1	0	0	0	1	1	0	1
23	0	0	0	1	28	1	0	0	0	0	1	0	1

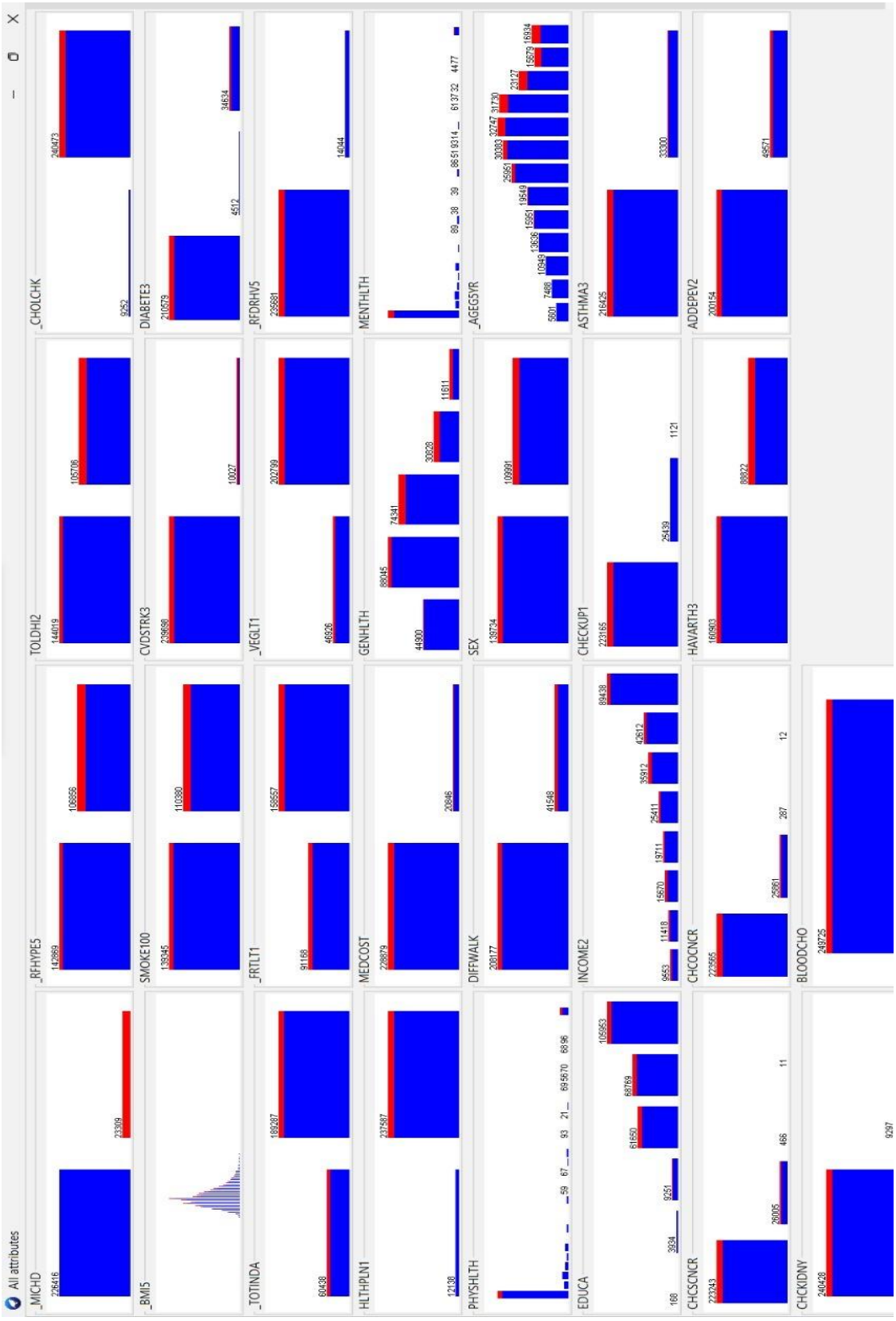
Şekil 21 Veri Setinin Son Hali-2 (Weka)

Weka paket programı başlangıçta tüm verileri numerik gördüğü için veriler nominal hale getirilmiştir. Weka'da veri ön işleme ekranı Şekil 22'de verilmiştir.



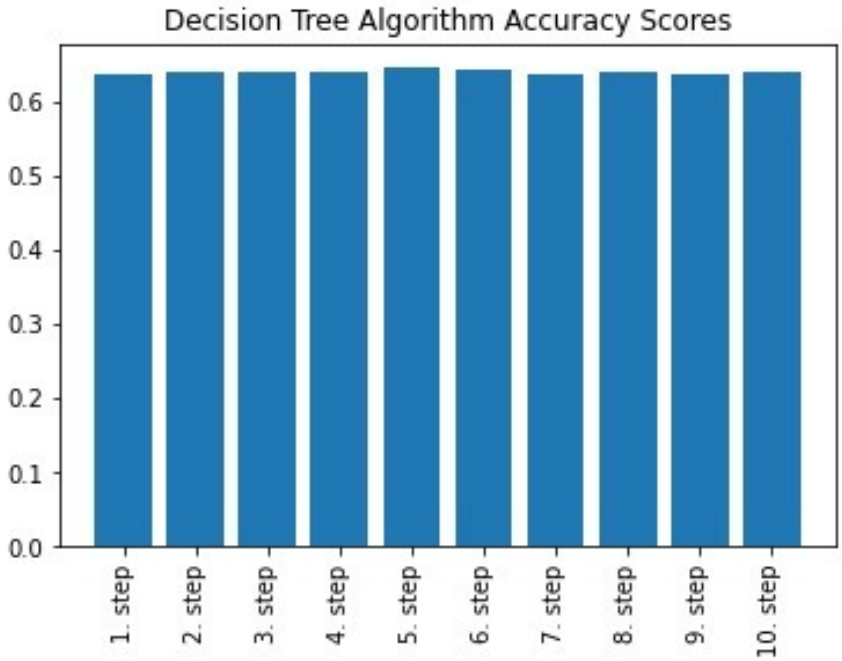
Şekil 22 Weka'da Veri Ön İşleme Ekranı

Tüm değişkenler için Weka paket programından elde edilen grafikler Şekil 23'de verilmiştir

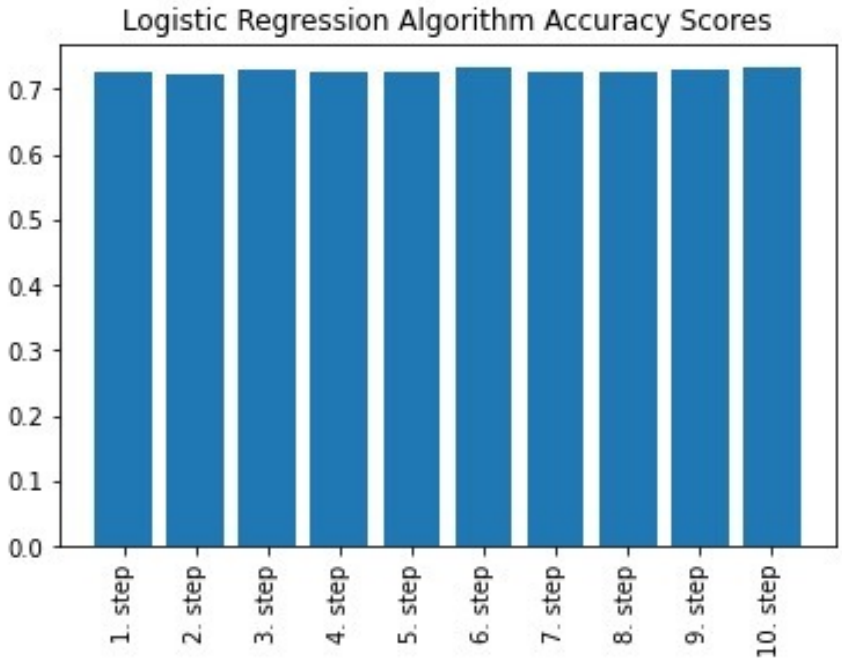


Şekil 23 Weka Paket Programından Elde Edilen Grafikler

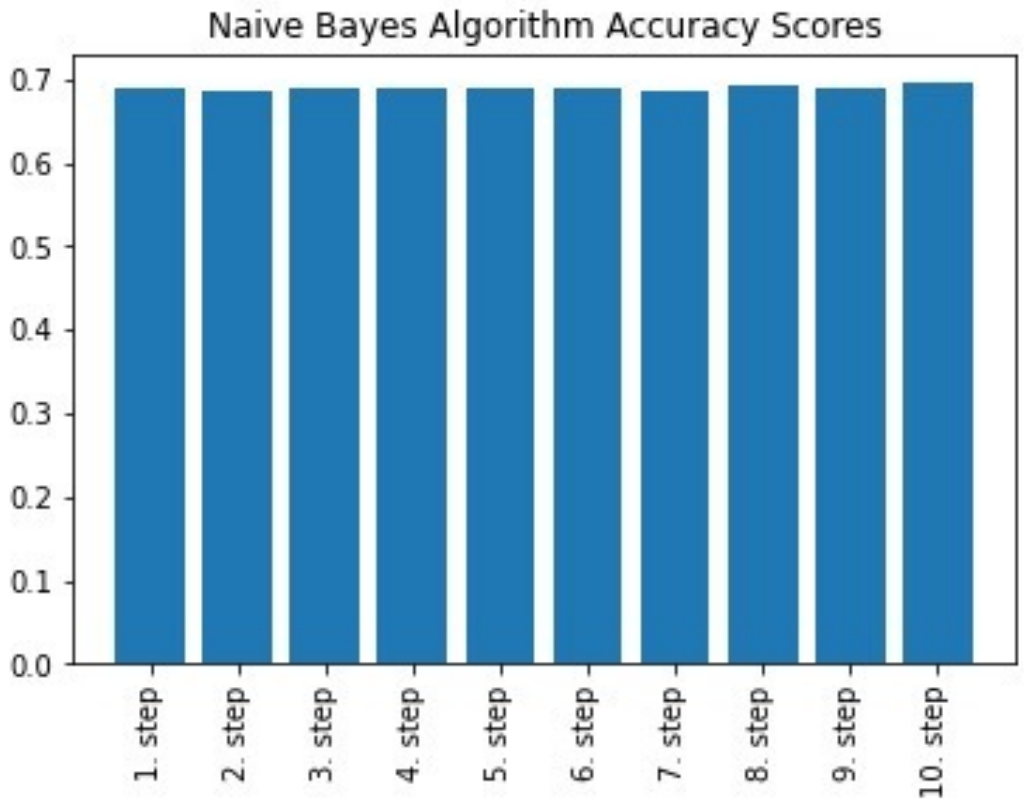
Bağımlı değişkende 1'lerin sayısı sıfırların sayısından çok daha fazla olduğu için 10 adımlı çapraz doğrulama yapılmıştır. Her bir algoritma için çapraz doğrulama süreci sonucunda elde edilen Şekil 24, 25, 26, 27, 28 ve 29'da verilmiştir.



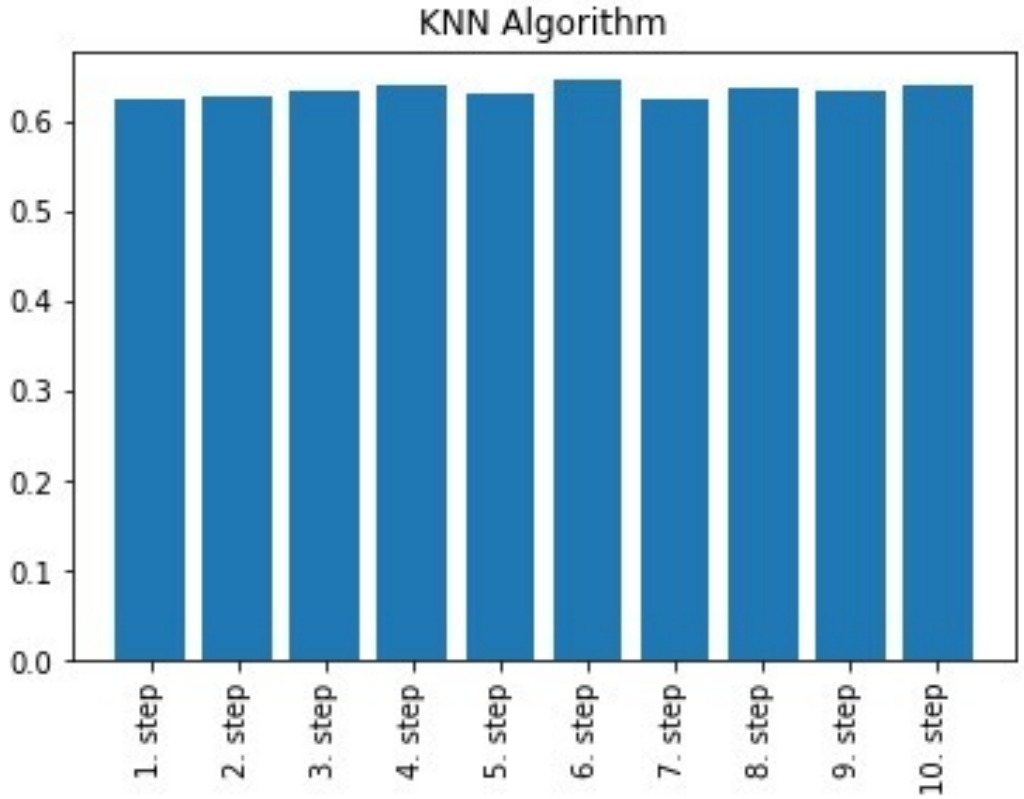
Şekil 24 Çapraz Doğrulama Süreci Sonucu-1



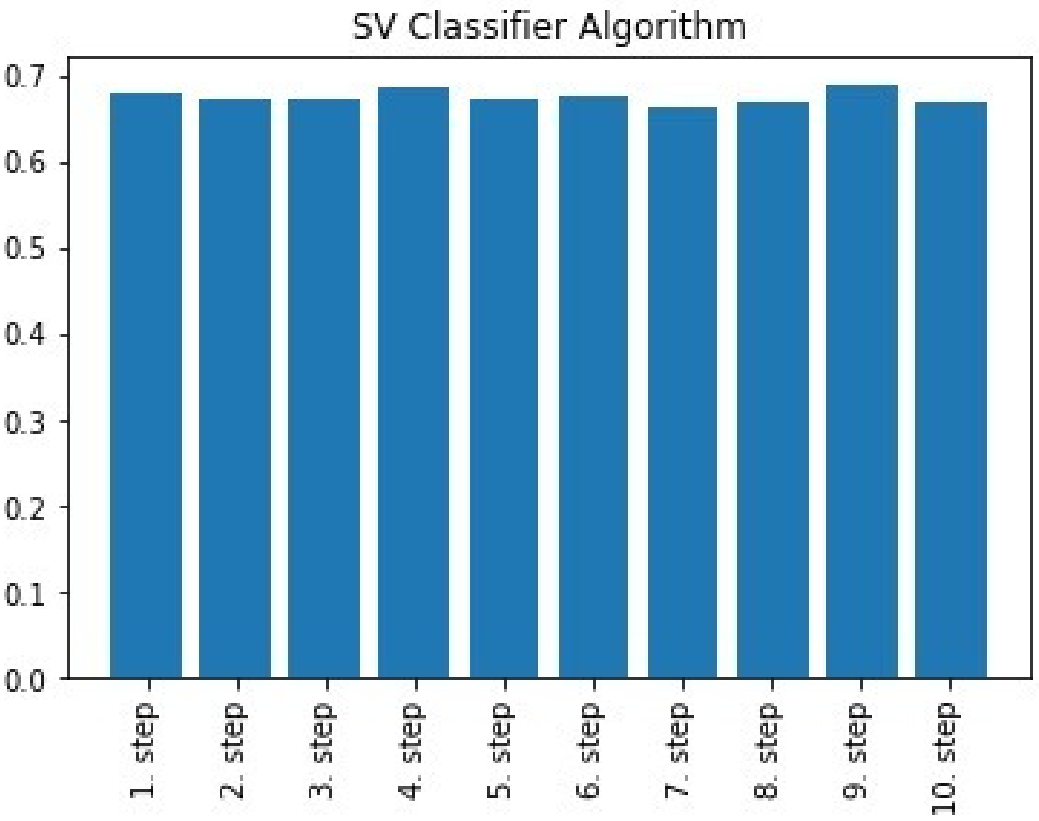
Şekil 25 Çapraz Doğrulama Süreci Sonucu-2



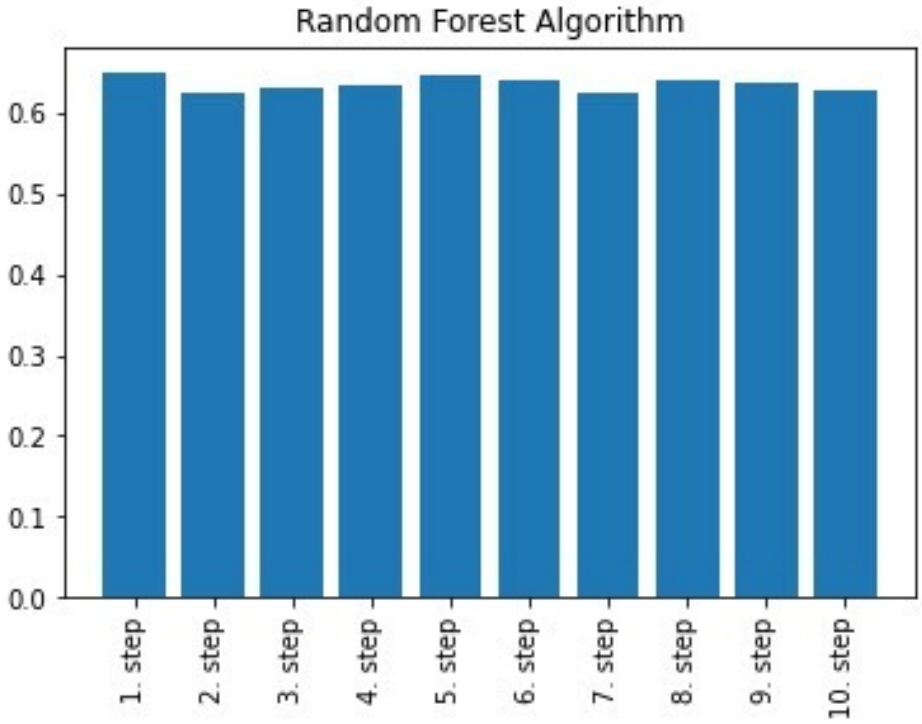
Şekil 26 Çapraz Doğrulama Süreci Sonucu-3



Şekil 27 Çapraz Doğrulama Süreci Sonucu-4



Şekil 28 Çapraz Doğrulama Süreci Sonucu-5



Şekil 29 Çapraz Doğrulama Süreci Sonucu-6

V. ANALİZ VE BULGULAR

Çalışmada kullanılan veri seti Amerika Birleşik Devletleri'nin Hastalık Kontrol ve Korunma Merkezleri (CDC) tarafından yürütülen, Davranışsal Risk Faktörü Gözetim Sistemi (BRFSS) anketinden elde edilmiştir. Veri anket üzerinden elde edildiği için öncelikle veri ön işleme sürecinden geçirilmiştir. Burada eksik gözlemler ile "cevap vermek istemiyorum" ve "bilmiyorum" cevaplarına sahip bireyler örneklemden düşürülmüş ardından kategorik değişkenler ikili değişken haline getirilmiştir. Bu bağlamda özellik seçimi sürecinden önce başlangıçta çalışmada kullanılan değişkenler Çizelge 2'de verilmiştir.

Çizelge 2 Kullanılan Değişkenler ve Tanımları

Değişken	**Tanım**
_MICHD	Daha önce koroner kalp hastalığı (KKH) veya miyokard enfarktüsü (MI) olduğunu bildirmiş olan katılımcılar için 1 diğerleri için 0 değerini alan ikili değişken
_RFHYPE5	Bir doktor, hemşire veya başka bir sağlık uzmanı tarafından yüksek tansiyonu olduğu söylenen yetişkinler için 1 diğerleri için 0 değerini alan ikili değişken
TOLDHI2	Bir doktor, hemşire veya başka bir sağlık uzmanı tarafından kan kolesterolünüzün yüksek olduğu söylenen kişiler için 1diğerleri için 0 değerini alan ikili değişken
_CHOLCHK	Kişi son 5 yıl içinde kolesterol kontrolü yaptırdıysa 1 aksi halde 0 değeri alan ikili değişken
SMOKE100	Kişi hayatı boyunca en az 100 sigara içtiyse 1 aksi takdirde 0 değerini alan ikili değişken [Not: 5 paket = 100 sigara]
_TOTINDA	Son 30 gün içinde normal işleri dışında fiziksel aktivite veya egzersiz yaptığını bildiren yetişkinler için 1 diğerleri için 0 değerini alan ikili değişken
_FRTLT1	Günde 1 veya daha fazla kez meyve tüketen bireyler için 1 diğerleri için 0 değerini alan ikili değişken
_VEGLT1	Günde 1 veya daha fazla kez sebze tüketen bireyler için 1 diğerleri için 0 değerini alan ikili değişken
_RFDRHV5	Ağır içiciler (haftada 14'ten fazla içki içen yetişkin erkekler ve haftada 7'den fazla içki içen yetişkin kadınlar) için 1 diğerleri için 0 değerini alan ikili değişken
HLTHPLN1	Sağlık sigortası, HMO'lar gibi ön ödemeli planlar veya Medicare veya Indian Health Service gibi hükümet planları dahil olmak üzere herhangi bir sağlık sigortası olanlar için 1 diğerleri için 0 değerini alan ikili değişken
MENTHLTH	Bireyin 30 gün içinde zihinsel sağlığının iyi olmadığı gün sayısı

Çizelge 2 Kullanılan Değişkenler ve Tanımları Devam

Değişken	Tanım
PHYSHLTH	Bireyin 30 gün içinde fiziksel sağlığının iyi olmadığı gün sayısı
DIFFWALK	Yürümekte zorluk çeken bireyler için 1 diğerleri için sıfır değerini alan ikili değişken
SEX	Kadınlar için 0 erkekler için 1 değerini alan ikili değişken
_AGEG5YR	On dört seviyeli yaş kategorisi
EDUCA	Bireyin en yüksek eğitim seviyesini temsil eden değişken
CHECKUP1	Bireyin son doktor kontrolünden sonra geçen süre
BLOODCHO	Kişi kan kontrolü yaptırdıysa 1 aksi takdirde 0 değerini alan ikili değişken
ASTHMA3	Astımı olan bireyler için 1 diğerleri için sıfır değerini alan ikili değişken
CHCSCNCR	Daha önce deri kanseri geçiren bireyler için 1 aksi takdirde 0 değerini alan ikili değişken
CHCOCNCR	Herhangi bir tür kanser geçiren bireyler için 1 diğerleri için 0 değerini alan ikili değişken
HAVARTH3	Bireyin bir çeşit artrit, romatoid artrit, gut, lupus veya fibromiyaljini varsa 1 aksi takdirde 0 değerini alan ikili değişken
ADDEPEV2	Bireyin depresyon, majör depresyon, distimi veya minör depresyon gibi depresif bir bozukluğu varsa 1 aksi halde 0 değerini alan ikili değişken
CHCKIDNY	Kişinin böbrek hastalığı varsa 1 aksi takdirde 0 değerini alan ikili değişken
DIABETE3	Birey diyabet hastalığına sahipse 1 aksi takdirde 0 değerini alan ikili değişken

Veri ön işleme süreci tamamlandıktan sonra özellik seçimi sürecine geçilmiştir. Özellik seçimi, makine öğrenimi modelleri kullanılan çalışmalarda açıklayıcı değişken sayısını bir diğer deyişle girdi sayısını azaltmak için kullanılan bir yöntemdir. Bu yöntemin kullanılma sebebi girdi sayısının fazla olması sebebiyle öğrenme sürecinin yavaş olması ya da yanıt (bağımlı değişken) değişkeniyle korelasyonsuz değişkenlerin modelin gücünü düşürmesi olabilir. Ki regresyona dayalı makine öğrenmesi algoritmalarında, bağımlı değişken ile ilgisiz değişkenlerin modele dahil edilmesi sonucunda ekonometrik analizde sıkça görülen "gereksiz değişkenin modele dahil edilmesi" problemi ortaya çıkabilir. Özellik seçimi tümdengelim ya da tümevarım yoluyla yapılabilir. Sıralı ileri seçim tümevarıma dayanır ki bu yöntemin başlangıcında herhangi bir girdi değişkenine sahip olmayan bir model kurulur ve sonrasında tüm

girdi değişkenleri sırasıyla modele dahil edilir. Her bir adımda model seçim kriterlerine bakılarak (AIC:$2k - 2\ln(L)$, düzeltilmiş R kare:$1 - \frac{(1-R^2)(n-1)}{n-k-1}$, eklenen değişkenin katsayısının t istatistiğinin olasılık değeri: $\frac{a}{(2-n-k)}$, hata kareler toplamı:$\sum u^2$) model performansına katkısı olan girdi değişkenleri modelde kalır diğerleri model dışı bırakılır sonuç olarak nihai model elde edilir. Sıralı geri seçim ise tümdengelime dayanır, başlangıçta tüm girdi değişkenlerinin bulunduğu bir model tahmin edilir ve sırasıyla her bir adımda istatistiksel olarak anlamsız olan değişkenler model dışı bırakılarak nihai model elde edilir.

Bu çalışmada bağımlı (yanıt değişkeni) değişken ikili bir değişken olduğu için özellik seçimi klasik doğrusal regresyon ile değil lojistik regresyon ile yapılmaktadır. Bağımlı değişkenin ikili olduğu durumlarda, doğrusal regresyon kullanılırsa hata teriminde ortaya çıkan değişen varyans sebebiyle katsayıların standart hataları yanlı tahmin edilir ve dolayısıyla t istatistikleri ve p-değerleri olduğundan büyük ya da küçük bulunur. Sıralı geriye seçim yapılırken aslında katsayısı istatistiksel olarak anlamlı olan bir değişkeni -yanlış tahmin edilen p-değerinden dolayı- modelden çıkartmak ya da katsayısı istatistiksel olarak anlamlı olmayan bir değişkeni -yanlış tahmin edilen p-değerinden dolayı- modelden tutmaktan kaçınmak için özellik seçiminde lojistik regresyon kullanılmıştır.

Geriye doğru eleme yöntemi kullanılarak yapılan özellik seçimi sonucunda %5 önem düzeyinde istatistiksel olarak anlamsız olan FRTL1, PYSHLTH, BMI5, EDUCA, HLTHPLN1, CHCOCNCR, CHECKUP1 ve CHCSCNCR değişkenleri modelden dışlanarak nihai model elde edilmiştir.

Kalp hastalıklarının tespitinde hangi makine öğrenim algoritmasının başarılı olduğunu tespit etmek amacıyla yapılan bu çalışmada nihai model sınıflandırma algoritmalarından, sağlık konusunda çalışılırken sıklıkla kullanılan K En Yakın Komşu, Lojistik Regresyon, Destek Vektör Makineleri, Karar Ağaçları, Rassal Ormanlar, Naive Bayes ve Yapay Sinir Ağları aracılığıyla analiz edilmiştir. Bu algoritmaların kalp hastalıklarını tespit etmedeki gücü başarı oranı, kesinlik skoru, duyarlılık skoru ve F1 skoru açısından kıyaslanmıştır. Başarı oranı, toplam doğru sınıflandırmaların, toplam gözlemlere oranı olarak tanımlanır:

$$BO = \frac{DPS + DNS}{TS} \quad \text{(Denklem 3)}$$

BO = Başarı Oranı

DPS = Doğru Pozitiflerin Sayısı

DNS = Doğru Negatiflerin Sayısı

TS = Toplam Gözlem Sayısı

Başarı oranı yükseldikçe modelin uyum iyiliğinin de arttığı söylenebilir. Kesinlik skoru, doğru pozitiflerin toplam pozitifler içindeki payı olarak tanımlanırken,

$$K = \frac{DPS}{TPS} \quad \text{(Denklem 4)}$$

K = Kesinlik

DPS = Doğru Pozitiflerin Sayısı

TPS = Toplam Pozitiflerin Sayısı

Hassasiyet skoru, doğru pozitiflerin doğru sınıflandırmalar içindeki payıdır.

$$H = \frac{DNS}{DPS + YNS} \quad \text{(Denklem 5)}$$

H = Hassasiyet

DNS = Doğru Negatiflerin Sayısı

DPS = Doğru Pozitiflerin Sayısı

YNS = Yanlış Negatiflerin Sayısı

F1 skoru ise, kesinlik ve hassasiyetin çarpımının 2 katının, kesinlik ve hassasiyet içindeki payı olarak tanımlanmaktadır.

$$F1 = \frac{2 \times K \times H}{K + H} \quad \text{(Denklem 6)}$$

F1 = F1 Skoru

K = Kesinlik

H = Hassasiyet

Tüm bu sınıflandırma kriterlerinin her bir algoritma için sonuçları Çizelge 3'de verilmiştir.

Çizelge 3 Algoritma Sonuçları

	Başarı oranı	**Kesinlik**	**Duyarlılık**	**F1 Skoru**
K En Yakın Komşu	0.8625	0.59	0.58	0.58
Lojistik Regresyon	0.9077	0.74	0.56	0.58
Destek Vektör Makineleri	0.9052	0.45	0.50	0.48
Karar Ağaçları	0.8725	0.60	0.58	0.59
Rassal Ormanlar	0.8922	0.63	0.56	0.57
Naive Bayes	0.8722	0.61	0.69	0.63
Yapay Sinir Ağları	0.9054	0.76	0.53	0.54

Modellerin başarılı tahmin oranlarına bakıldığında en yüksek başarıya sahip modellerin Lojistik Regresyon, Destek Vektör Makineleri ve Yapay Sinir Ağları olduğu görülmektedir. Modellerin kesinlik skorlarına bakıldığında ise en yüksek başarıyla tahmin yapan algoritmaların Lojistik Regresyon ve Yapay Sinir Ağları olduğu görülmektedir. Duyarlılık skoru açısından bakıldığında ise en iyi model Naive Bayes'tir. F1 skorlarına bakıldığında ise en başarılı modeller Karar Ağaçları ve Naive Bayes'tir.

Çalışma açısından bu model başarı kriterlerine bakıldığında, uç gözlemlere sahip veri kümelerinde başarı oranı, kesinlik ve duyarlılığına bakılması doğru değildir ancak bu çalışmada veri ön işleme sürecinde değişkenler ikili değişken haline getirildiği için veri setinde uç gözlemler bulunmamaktadır. Bu sebeple başarı oranı, kesinlik ve duyarlılığa bakılarak en başarılı modellerin Lojistik Regresyon ve Yapay Sinir Ağları modelleri olduğuna karar verilir.

VI. SONUÇ

Dünya yer alan birçok ülkede kalp rahatsızlıkları gerek kadınlar gerekse de erkekler arasında çok yaygın görülmektedir. Bu sebep ile bireyler kalpte meydana gelecek risk etmenlerini dikkate almalıdır. Bazı etmenlerini yaşam tarzı etmenleri, bazı etmenler genetik bir rol oynamaktadır ve bu durumlar kalpte ciddi derecede etki etmektedir. Bütün bunlara bakıldığında çalışmada makine öğrenme yöntemleri ele alınmış ve incelenmiştir. Dünya'da kullanımı git gide artış gösteren bu yöntemi birçok alt başlığı yer aldığı gibi kullanılacak alana göre başlıkların avantajları değişkenlik göstermektedir. Araştırma içerisinde destek vektör makineleri, rastgele orman algoritması, yapay sinir ağları, k-NN (En Yakın Komşu), lojistik regresyon, karar ağaçları ve Naive Bayes yöntemleri karşılaştırılmıştır.

Çalışmanın sonucunda kalp rahatsızlıklarını tespit etmede en başarılı makine öğrenim algoritmasının başarı oranlarına (accuracy rate) göre lojistik regresyon, destek vektör makinaları ve yapay sinir ağları; kesinlik skorlarına göre lojistik regresyon ve yapay sinir ağları; duyarlılık skorlarına göre Naive Bayes ve F1 skorlarına göre karar ağaçları ve Naive Bayes olduğu görülmüştür.

Uç gözlemlere sahip veri kümelerinde modelleri kıyaslamak amacıyla başarı oranı, kesinlik ve duyarlılık skorlarına bakmak doğru değildir ancak bu çalışmada veri ön işleme sürecinde tüm değişkenler ikili değişken haline getirildiği için veri setinde uç gözlemler bulunmamaktadır. Dolayısıyla, başarı oranı, kesinlik ve duyarlılık skorlarına bakılarak kalp hastalıklarını tespit etmede en başarılı makine öğrenimi algoritmalarının Lojistik Regresyon ve Yapay Sinir Ağları modelleri olduğuna karar verilir.

Çalışmada kullanılan bağımlı değişken ikili değişken olduğu için bu çalışmada sınıflandırma algoritmaları kullanılmıştır, kalp rahatsızlığını temsil edilebilecek sürekli bir bağımlı değişken elde edilebilirse çalışmanın yalnızca sınıflandırma algoritmalarına sahip olması kısıtı ortadan kaldırılabilir.

Bu çalışmayı diğer çalışmalardan ayıran temel nokta model kıyaslamasını RMSE (Kök Ortalama Kare Hata) kriteri üzerinden değil model seçim kriterleri üzerinden yapılmasıdır.

Makine öğrenmesi algoritmalarının daha büyük veri setleriyle daha iyi çalıştığı göz önünde bulundurulduğunda daha geniş bir veri setiyle çalışılarak çalışma daha ileri noktalara taşınabilir. Ayrıca, eğer kalp rahatsızlığına sahip kişilerin kalplerine ait görüntüler elde edilebilirse ilerlenen çalışmalarda görüntü işleme algoritmaları kullanılarak daha başarılı sonuçlar elde edilebilir.

VII. KAYNAKÇA

KİTAPLAR

ALAPYDİN, E. (2010). **Introduction to Machine Learning**, 2nd ed. Cambridge Massachusetts, MIT Press.

BALDİ, P., & BRUNAK, S. (2001). **Bioinformatics: the machine learning approach**. MIT press.

BRAMER, M. (2013). **Principles of Data Mining**. 2nd ed. Springer.

CORD, M., & CUNNİNGHAM, P. (Eds.). (2008). **Machine learning techniques for multimedia: case studies on organization and retrieval**. Springer Science & Business Media.

DASH, M. & LİU, H. (1997). **Feature Selection for Classification**, Intelligent Data Analysis, Elsevier.

GARETH J. (2013). **Introduction to Statistical Learning**. New York, Springer.

JENSON, P. (2001). **Bayesian Networks And Decision Graph**s, Springer, New-york, USA.

LOPEZ, E.O., & BALLARD, B.D. (2021). **Heart Disases. StatPearls**. Hazine Adası (FL): StatPearls Yayıncılık.

MACKENZİE, J. (2005). **Heart Disases**. Oxford Medikal Yayınları.

PROVOST, F. (2000). Distributed data mining: Scaling up and beyond. **Advances in distributed and parallel knowledge discovery**, Foster Provost New York University.

SHALEV-SHWARTZ, S., & BEN-DAVİD, S. (2014). **Understanding machine learning: From theory to algorithms**. Cambridge university press.

TETTAMANZİ, A. & TOMASSİNİ, M. (2001). **Soft Computing**. Springer-Verlag.

WİLSON, R. (1999). **The MIT Encyclopaedia of Cognitive Sciences**. MIT Press.

MAKALELER

ABACI, A. (2011). Kardiyovasküler risk faktörlerinin ülkemizdeki durumu. **Türk Kardiyol Dern Arş-Arch Turk Soc Cardiol**, *39*(4), 1-5.

AHER, S., LOBO, M. (2011). Data Mining in Educational System using WEKA, **International Conference on Emerging Technology Trends (ICETT)**

BENJAMİN E.J., VİRANİ S.S., CALLAWAY C.W., Et al., (2018). American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart Disease and Stroke Statistics. A Report From the American Heart Association. **Circulation**. 137(12).

BULUT, F. (2016). Determining Heart Attack Risk Ration Through AdaBoost/AdaBoost ile Kalp Krizi Risk Tespiti. **Celal Bayar University Journal of Science,** *12*(3), 459-472.

CARTER, R. J., DUBCHAK, I., & HOLBROOK, S. R. (2001). A computational approach to identify genes for functional RNAs in genomic sequences. **Nucleic acids research,** 29(19), 3928-3938.

COŞAR, M., & DENİZ, E. (2021). Makine Öğrenimi Algoritmaları Kullanarak Kalp Hastalıklarının Tespit Edilmesi. **Avrupa Bilim ve Teknoloji Dergisi**, Özel Sayı 28, S. 1112-1116.

ÇALIŞ, A., KAYAPINAR, S., & ÇETİNYOKUŞ, T. (2014). Veri Madenciliğinde Karar Ağacı Algoritmaları İle Bilgisayar Ve İnternet Güvenliği Üzerine Bir Uygulama. **Endüstri Mühendisliği**, *25*(3), 2-19.

DİEHM, C., LANGE, S., DARİUS, H., PİTTROW, D., VON STRİTZKY, B., TEPOHL, G., ... & TRAMPİSCH, H. J. (2006). Association of low ankle brachial index with high mortality in primary care. **European heart journal,** *27*(14), 1743-1749.

DÜLEK, H., TUZCULAR VURAL, E. Z., & GÖNENÇ, I. (2018). Kardiyovasküler hastalıklarda risk faktörleri. **The Journal of Turkish Family Physician**, *9*(2), 53-58.

FARLEY A., MCLAFFERTY E., & HENDRY C. (2012). The Cardiovascular System. 31 Ekim - 6 Kasım 2012. **Nurs Stand.** 27(9):35-9 s.

GÖKTAŞ, M.E., YAĞANOĞLU, M. (2020). Veri Bilimi Uygulamalarının Hastalık Teşhisinde Kullanılması: Kalp Krizi Örneği. **Bilişim Sistemleri ve Yönetim Araştırmaları Dergisi** 2(2), 26-32.

GUYON, I. & A. ELİSSEEFF, (2003). "An introduction to variable and feature selection," **Journal of Machine Learning Research**, vol. 3, pp. 1157-1182.

GÜNDOĞDU, S. (2020). Kalp Hastalık Risk Tahmini İçin Python Aracılığıyla Sınıflandırıcı Algoritmalarının Performans Değerlendirmesi. **Dokuz Eylül Üniversitesi Fen ve Mühendislik Dergisi.** 23(69), 1005-1013 s, 2021.

HİRA, Z., & GİLLİES, D. (2015). A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. **Hindawi Publishing Corporation Advances in Bioinformatics** Volume 2015, Article ID 198363, 13 pages.

HOSSAİN, M. R., OO, A. M. T., & ALİ, A. B. M. S. (2013). The combined effect of applying feature selection and parameter optimization on machine learning techniques for solar Power prediction. **American Journal of Energy Research**, *1*(1), 7-16.

KARAKOÇ KUMSAR, A., TAŞKIN YILMAZ, F. (2017). Kardiyovasküler Hastalıklar Risk Faktörlerinden Korunmada Hemşirenin Rolü. **Online Türk Sağlık Bilimleri Dergisi** 2017, 2(4,) 18-27.

KUMAR, V., & MİNZ, S. (2014). Feature selection: a literature review. **SmartCR**, *4*(3), 211-229.

KÜLTÜRSAY, H. (2011). Kardiyovasküler hastalık riski hesaplama yöntemleri. **Türk Kardiyol Dern Arş-Arch Turk Soc Cardiol,** *39*(4), 6-13.

LAKSHMİ, K. R., KRİSHNA, M. V., & KUMAR, S. P. (2013). Performance comparison of data mining techniques for predicting of heart disease survivability. **International Journal of Scientific and Research Publications,** *3*(6), 1-10.

LU, L., & ZHU, Z. (2014). Prediction model for eating property of indica rice. **Journal of Food Quality**, 37(4), 274-280.

MAHESH, B. (2018). Machine Learning Algorithms A Review. **International Journal of Science and Research** (IJSR) ISSN: 2319-7064.

MOLİNA, L. C., BELANCHE, L., & NEBOT, À. (2002, December). Feature selection algorithms: A survey and experimental evaluation. **IEEE International Conference on Data Mining**, (pp. 306-313).

NİSSEN, S. (2003). Implementation of a fast artificial neural network library (fann). *Report,* Department of Computer Science University of Copenhagen ***(DIKU),*** *31*(29), 26.

OKCU, Z., KELEŞ, F. (2011). Kalp-Damar Hastalıkları ve Antioksidanlar. **Atatürk Üniversitesi Ziraat Fakültesi Dergisi**. 40(1): 153-160.

ONAT, A., UĞUR, M., TUNCER, M., AYHAN, E., KAYA, Z., KÜÇÜKDURMAZ, Z., ... & KAYA, H. (2009). Age at death in the Turkish Adult Risk Factor Study: temporal trend and regional distribution at 56,700 person-years' follow-up. **Turk Kardiyol Dern Ars**, *37*(3), 155-60.

OZCAN, İ., TASAR, B., TATAR, A. B., & YAKUT, O. (2019). Destek Vektör Makinasi Algoritması İle Kalp Hastalıklarının Tahmini**. Computer Science,** *4*(2), 74-79.

ÖZMEN, Ö., KHDR, A., & AVCI, E. (2018). Sınıflandırıcıların Kalp Hastalığı Verileri Üzerine Performans Karşılaştırması. **Fırat Üniversitesi Müh. Bil. Dergisi**, 30(3).

PANDA, M., & PATRA, M. R. (2007). Network intrusion detection using naive bayes. **International journal of computer science and network security**, *7*(12), 258-263.

PONRAJ, T.C., & RAMESH, S. S. (2020). Minimizing Influence Of Rumours On Social Networks Using Machine Learning Algorithms And Analysis. **Journal Of Mechanics Of Continua And Mathematical Sciences.** 15(5).

TÜRKMEN E, BADIR A, & ERGÜN A. (2012). Koroner arter hastalıkları risk faktörleri: Primer ve sekonder korumada hemşirelerin rolü. **Acıbadem Üniversitesi Sağlık Bilimleri Dergisi,** 3(4):223-31.

TÜRKMEN, E., & GÜVEN GS. (2010). Kardiyovasküler hastalıklardan primer korunma esasları. **Hacettepe Tıp Dergisi**, 41(3):179-85.

WİTTEN, I. H., FRANK, E., HALL, M. A., PAL, C. J., & DATA, M. (2005). Practical Machine Learning Tools And Techniques. **Elsevier.**

XİE Z, NİKOLAYEVA O, LUO J, Lİ D. (2019). Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. **Prev Chronic Dis**, (16): DOI: http://dx.doi.org/10.5888/pcd16.190109external icon.

ZAKARİA, M., AL-SHEBANY, M., SARHAN, S. (2014). Artificial Neural Network : A Brief Overview. Mabrouka AL-Shebany et al. **Int. Journal of Engineering Research and Applications**, 4(2).

ELEKTRONİK KAYNAKLAR

URL-1 Kardiyavasküler Atlası, WHO, https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1 (Erişim Tarihi: 22.03.2022)

URL-2 Ölüm Nedeni İstatistikleri, TÜİK, https://data.tuik.gov.tr/Bulten/Index?p=Olum-ve-Olum-Nedeni-Istatistikleri-2019-33710 (Erişim Tarihi: 22.03.2022)

URL-3 Kardiyovasküler Hastalıklar, WHO, https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (Erişim Tarihi: 08.04.2022)

URL-4 Centers for Disease Control and Prevention https://www.cdc.gov/heartdisease/index.htm (Erişim Tarihi: 08.04.2022)

URL-5 CSC Infographic Big Data [online]. URL:http://assets1.csc.com/insights/downloads/CSC_Infographic_Big_Data.pdf Accessed 26 February 2017 Erişim Tarihi: 10.04.2022)

URL-6 Business Dictionary. http://www.businessdictionary.com/definition/regression.html (Erişim Tarihi: 10.04.2022)

URL-7 MathWorks https://www.mathworks.com/discovery/feature-extraction.html (Erişim Tarihi: 17.04.2022)

URL-8 Microsoft Azure https://azure.microsoft.com/tr-tr/overview/machine-learning-algorithms/#overview (Erişim Tarihi: 17.04.2022)

URL-9 A Basic Introduction To Neural Networks https://users.cs.duke.edu/~brd/Teaching/Previous/AI/Lectures/NN/neural.html M (Erişim Tarihi: 18.04.2022)

URL-10 IBM Neural Networks https://www.ibm.com/cloud/learn/neuralnetworks#:~:text=Neural%20networks%2C%20also%20known%20as,neurons%20signal%20to%20one%20another. (Erişim Tarihi: 20.04.2022)

URL-11 Michael Luk. K-Nearest Neighbour and Dynamic Time Wrapping. Devices using DTW and KNN. https://sflscientific.com/case-studies/2016/6/4/time-series-analysis-fitbitusing-dtw-and-knn (Erişim Tarihi: 25.04.2022)

URL-12 Machine Learning Classification Naive Bayes https://medium.com/@ekrem.hatipoglu/machine-learning-classification-naive-bayes-part-11-4a10cd3452b4 (Erişim Tarihi: 25.04.2022)

TEZLER

CARREİRA-PERPİÑÁN, M. (1997). **A Review of Dimension Reduction Techniques. Technical Report CS–96–09** Dept. of Computer Science, University of Sheffield.

GÖRGÜN, M. (2020). **Makine Öğrenmesi Yöntemleri İle Kalp Hastalığının Teşhis Edilmesi**, Yüksek Lisans Tezi, İstanbul Aydın Üniversitesi.

EKLER

#MAKİNE ÖĞRENMESİ ALGORİTMALARIYLA KALP HASTALIKLARININ TESPİT EDİLMESİNE YÖNELİK PERFORMANS ANALİZİ

```
#Python
import statsmodels.api as sm
import pandas as pd
import numpy as np
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
import numpy as np
import pandas as pd
from sklearn.preprocessing import LabelEncoder,OneHotEncoder
dataset = pd.read_csv('cleandata.csv')

X = dataset.iloc[:,2:]
y = dataset.iloc[:,1:2]

logit_model=sm.Logit(y,X)
result=logit_model.fit()
print(result.summary2())

model = sm.OLS(y,X)
results = model.fit()
print(results.summary2())

dataset = dataset.drop(['_FRTLT1'],axis=1)

dataset = dataset.drop(['PHYSHLTH'],axis=1)
```

```
dataset = dataset.drop(['_BMI5'],axis=1)

dataset = dataset.drop(['EDUCA'],axis=1)

dataset = dataset.drop(['HLTHPLN1'],axis=1)

dataset = dataset.drop(['MENTHLTH'],axis=1)

dataset = dataset.drop(['CHCOCNCR'],axis=1)

dataset = dataset.drop(['CHECKUP1'],axis=1)

dataset = dataset.drop(['CHCSCNCR'],axis=1)

X = dataset.iloc[:,2:]
y = dataset.iloc[:,1:2]

from sklearn.model_selection import train_test_split

x_train, x_test,y_train,y_test = train_test_split(X,y,test_size=0.33, random_state=0)

y=np.reshape(y, (299,1))
y=pd.DataFrame(y)


from sklearn.preprocessing import MinMaxScaler

sc=MinMaxScaler()
```

```
X = sc.fit_transform(X)
y= sc.fit_transform(y)

from sklearn.preprocessing import LabelEncoder

le = LabelEncoder()
X[:,1] = le.fit_transform(X[:,1])

le2 = LabelEncoder()
X[:,2] = le2.fit_transform(X[:,2])

from sklearn.preprocessing import OneHotEncoder
ohe = OneHotEncoder(categorical_features=[1])
X=ohe.fit_transform(X).toarray()
X = X[:,1:]

from sklearn.model_selection import train_test_split
x_train, x_test,y_train,y_test = train_test_split(X,y,test_size=0.25,random_state=2)

df=pd.DataFrame(y_test)
df.to_excel("y_test.xlsx")
df1=pd.DataFrame(y_pred)
df1.to_excel("y_pred1.xlsx")

import keras
from keras.models import Sequential
from keras.layers import Dense

classifier = Sequential()
```

```
classifier.add(Dense(4, init = 'uniform', activation = 'relu' , input_dim =20))
#hidden layer

classifier.add(Dense(6, init = 'uniform', activation = 'relu'))



classifier.add(Dense(1, init = 'uniform', activation = 'relu'))

classifier.compile(optimizer = 'adamax', loss =  'mean_squared_error' , metrics = ['accuracy'] )

classifier.fit(x_train, y_train, epochs=10)

y_pred = classifier.predict(x_test)

y_pred[y_pred<0.5] = 0
y_pred[y_pred>0.5] = 1

from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report

cm = confusion_matrix(y_test,y_pred)
print('KNN')

succ = cm[0][0]+cm[1][1]
succ_knn = succ/len(y_test)
print(succ_knn)
cls_report= classification_report(y_test,y_pred)
```

ÖZGEÇMİŞ

Elif Çil

Profesyonel Deneyim

2017-2021: Nazmi Arıkan Fen Bilimleri Okulları (Robotik Yazılım Ve Kodlama Öğretmeni)

2021-2022 Denizatı Okulları (Robotik Yazılım Ve Kodlama Öğretmeni)

2022-Halen : Turkcell Global Bilgi (Uygulama Servis Operasyonları Uzmanı)

Eğitim Bilgileri

İstanbul Aydın Üniversitesi – Bilgisayar Mühendisliği (Yüksek Lisans) / 2019 –2022

Burs : %75 Onur Öğrencisi Bursu

Not Ortalaması – 3,02/ 4

İstanbul Aydın Üniversitesi – Bilgisayar Ve Öğretim Teknolojileri Öğretmenliği / 2012– 2016

Burs : %100 Ösym Bursu : 3,00 / 4

Pendik Türk Telekom Anadolu Teknik Lisesi – Bilişim Teknolojileri – Web Tasarım / 2008 – 2012