# Çevresel Faktörlerin Semen Kalitesi Üzerindeki Etkisinden Yola Çıkarak Tahminleme Yapılması

Fatma Selen AKAR, Hande ÇADIR
Ege Üniversitesi, Moleküler Biyoloji Bölümü
İzmir, Türkiye

## ÖZET

İnfertilite sorunu son yıllarda önemli bir sorun haline gelmiştir. Semen analizi, erkeklerin doğurganlık potansiyelini değerlendirmek için yapılan bir testtir. Yaşam alışkanlıkları ve sağlık durumunun semen kalitesini etkilediği birçok araştırmada görülmüştür.

Bu nedenle, bu çalışmada UCI Machine Learning Repository'de bulunan, 'Fertility Data Set'inden yararlanarak; mevsim, yaş, sigara ve alkol tüketimi, çocukluk hastalığı, travma, günlük oturma saatleri ve cerrahi müdahale gibi özniteliklerin sperm kalitesi üzerindeki etkisi üzerinden Support Vector Classifier, Karar Ağacı, Lojistik Regresyon, K-en yakın komşu, Naive Bayes ve XGBoost algoritmaları ile tahminleme yapmak amaçlanmıştır. Modellerin doğruluğu, kesinlik, duyarlılık, doğruluk ve F-ölçütü ile kontrol edilmiş, K-katlı cross validation kullanılmıştır.

Sınıflandırma sonucuna göre, %84 ile en yüksek F1 değerine sahip makine öğrenmesi algoritması Random Forest olurken bunu, %82 değeri ile XGBoost, Lojistik Regresyon, SVC ve kNN sınıflandırıcıların takip etmiştir. Naive Bayes %80 ve Karar Ağacı algoritmalarının %78 tahmin değeri ile bu veri seti için diğer algoritmalara göre daha az başarılı tahminde bulunduğu görülmüştür.

Semen kalitesinin tahminlenmesi açısından bu çalışma, sonraki çalışmalara öncü olarak tercih edilen algoritmanın geliştirilmesine ve iyileştirilmesine yardımcı olacaktır.

**Anahtar Kelimeler:** İnfertilite, Makine öğrenmesi, Decision Tree, Fertility, Logistic Regression, Naives Bayes, XGBoost, SVC, Random Forest

# 1.GİRİŞ

İnfertilite, dünya çapında milyonlarca insanın yaşadığı bir sağlık sorunudur. Dünya Sağlık Örgütü (WHO), infertiliteyi, 12 ay veya daha fazla düzenli korunmasız cinsel ilişkiden sonra gebelik elde edilememesi halinde erkek veya kadın üreme sisteminin bir hastalığı olarak tanımlamıştır ve dünya çapında bu sorunla karşılaşan 48 milyon çift ile 186 milyon kişi bulunduğunu açıklamıştır.

İnfertilite vakalarının yaklaşık %30'u yalnızca erkeğe bağlı sorunlardan oluşur. Erkeklerde infertilite nedenleri arasında hormonal bozukluklar, yaşam tarzı, fiziksel, psikolojik ve cinsel sorunlar, kromozomal anormallikler ve tek gen kusurları gibi çeşitli nedenler bulunur (Babakhanzadeh et al., 2020).

Spermin hareketi, canlılığı ve morfolojisi gibi faktörler hakkında bilgi almak için sperm analizi yapılır. Mesleki durum ve sağlıksız hayat tarzının sperm kalitesi üzerinde negatif etkisi bulunmaktadır (Wang et al., 2013). Danimarka'da yapılan bir çalışma, kullanılan alkol miktarının, sperm yoğunluğunu ve sayısını azalttığını göstermektedir (Virtanen et al., 2017).

İnfertilite üzerinde etkisi olan bir diğer etmen ise ateştir. Ateşin semen kalitesine etkisi ateşin süresine ve yüksekliğine göre değişkenlik göstermektedir. Ateşin varlığı ile sperm yoğunluğunda azalma, küçük başlı sperm sayısında artış ve anormal sperm formlarının üretiminde artış görülmüştür. Sperm kalitesinin eski haline dönmesi; ateşe ve ateşin iyileşme süresine bağlıdır. Hastaların yaklaşık dört ile beş hafta sonra sperm kalitesinde düzelme meydana gelmektedir (Andrade-Rocha, 2013).

Mevsimin sperm kalitesi üzerindeki etkisini inceleyen bir çalışmada ise, kış ve sonbahar aylarında sperm sayısında ve birtakım değişkenlerde artış olduğunu gözlemlenmiştir. Kafkas ve İsrail kökenli gençlerde yapılan araştırmada sperm konsantrasyonunun en fazla ocak ayında olduğu görülürken, en az eylül ayında görüldüğü tespit edilmiştir. Spermin morfolojisi incelendiğinde ise normal yüzdesinin en yüksek mart ayında, en düşük eylül ayında olduğu görülmüştür. Mevsimin, örnek hacmine ve hareketli sperm yüzdesinde bir etkisinin olmadığı görülürken, sonbahar ve yaza kıyasla sperm yoğunluğunun kış ve ilkbaharda daha fazla olduğu bulunmuştur (Yogev et al., 2004).

Bu çalışmada, çevresel faktörlerin semen kalitesi üzerindeki etkisinden yola çıkarak tahminleme yapılması amaçlanmıştır. Bu nedenle, son yıllarda oldukça fazla tercih edilen makine öğrenmesinden yararlanılmıştır.

Makine ögrenmesi algoritmaları olarak; Support Vector Classifier, Naive Bayes, Random Forest, Decision Tree, Logistic Regression, ve XGBoost kullanılmıştır.

## 2.ÖNCEKİ ÇALIŞMALAR

Günümüzdeki ciddi sağlık sorunlarından biri olan infertilitedeki artış küresel olarak yaklaşık 50 milyon çift doğurganlıkla ilgili sorunlar yaşamasına neden olmakta ve bu sorunların %30'u yalnızca erkek faktörlü infertiliteden kaynaklanmaktadır (You et al., 2021). İnfertilite sorunu yaşayan erkeklerin yaklaşık %50'sinde ise sorunun nedenin belirlenemediği görülmüştür. Erkeklerde infertilite kontrolü için altın standart olan semen analizinde, sperm sayısı, hareketliliği, fruktoz seviyesi ve pH gibi faktörlere bakılmaktadır.

Makine Öğrenimi, yapay zekanın gerçek dünya veri kümelerinden kalıpları bulmasına olanak sağlamakta ve veriye dayalı modelleme yeteneklerine yol açmıştır. Yapılan bir çalışmaya göre karar ağaçları, Gauss Naive Bayes sınıflandırıcısı ve lojistik regresyonun tahminlemesinin karşılaştırmasında lojistik regresyonun %88'lik umut verici bir doğruluk gösterdiği, bu modeller üzerinde bagging ensemble method kullandığında ise, en fazla gelişmenin %78'den %88'e karar ağacı algoritmasında olduğu görülmüştür (Dash & Ray, 2020).

Yapay zeka yöntemleri ile seminal kaliteyi tahmin etmeye yönelik başka çalışma örneklerine bakıldığında David Gi ve arkadaşlarının üç sınıflandırma yöntemini kullanarak seminal kalite tahminindeki doğruluğu değerlendirmek için karar ağaçları, çok katmanlı perceptron (tek katmanlı bir sinir ağı) (MLP) ve destek vektör makineleri (SVM) kullanarak yaptığı çalışmada daha iyi performans elde etmek için MLP kullandığı görülmüştür.

Wong ve arkadaşlarının yapmış olduğu bir çalışmada ise, seminal kalite tahmininde dengesiz sınıf öğrenme probleminin üstesinden gelmek için bir kümeleme tabanlı karar ormanları yöntemleri önermiştir (Bidgoli et al., n.d.). UCI fertility veri seti ile yapılan bir deep learning araştırmasında ise %89 ile başarılı sonuçlar elde edildiği görülmüştür (Benli et al., 2019).

İnfertilite haricinde de çeşitli hastalıklar üzerinde makine öğrenmesinin kullanıldığı, yakın tarihlerde yayınlanmış

olan çalışmalar da incelenmiştir. Bir araştırma çalışmasında, hastalıkların tahminlemesi için 4920 hastanın örnek verisi ve 41 hastalık teşhisi konulan kayıtlar analiz için seçilmiştir. Karar ağacı, random forest ve Naïve Bayes algoritmaları kullanılarak yapılan çalışma sonucuna göre bütün algoritmalar %95'e varan aynı doğruluğu ve performansı sağladığı görülmüştür (Grampurohit Sneha & Sagarnal Chetan, 2020).

Başka ciddi bir sağlık sorunu olan yağlı karaciğer hastalığı üzerine yapılan bir çalışmada, rastgele orman modeli, tahmine dayalı sınıflandırma modelleri arasında diğer modellere göre daha iyi performans gösterdiği görülmüş fakat İslam ve arkadaşlarının yapmış olduğu benzer bir çalışmada yukarıdaki sonucun aksine lojistik regresyon algoritmasının, diğer tüm makine öğrenimi algoritmalarına göre daha iyi sonuç (doğruluk, hassasiyet ve özgüllük) sağladığı görülmüştür (Wu et al., 2019) .

Kalp hastalıkları tahminlemesi için 14 temel özniteliğe sahip bir veri seti ile yapılan çalışmada K-en yakın komşu, Naïve Bayes ve Random forest algoritmaları (Shah et al., 2020), Kaggle 'dan alınan 12 özniteliğe sahip kardiyovasküler hastalık veri seti kullanılarak yapılan başka bir çalışmada ise karar Ağacı algoritması %73 doğruluk oranıyla en iyi tahminleri yaptıkları görülmüştür (Princy R.Jane Preetha et al., 2020).

Kronik obstrüktif akciğer hastalığındaki (KOAH) alevlenmelerin erken tespiti ve sonraki triyaj için makine öğrenimine dayalı bir çalışmanın sonucuna göre, en iyi performans gösteren iki algoritmanın, gradient boosting ve lojistik regresyon olduğu görülmüş fakat eksiklerden kaynaklı modellerin geliştirilmesi gerektiği fikrine varılmıştır (Swaminathan et al., 2017).

Böbrek Hastalığının prognoz oranını denetimli sınıflandırması için yapılan bir çalışmada lojistik regresyon ve random forest gibi dört mükemmel makine öğrenme tekniği kullanılarak başarılı sonuçlar elde edilmiştir. Bu modeller arasında lojistik regresyon %100 doğruluk oranını ile en başarılı algoritma olmuş, KNN algoritması diğerlerine göre daha yüksek hata oranına ve doğruluk değerine sahip olduğu görülmüştür. (Javed Mehedi Shamrat et al., 2020).

## 3.DENEYSEL ÇALIŞMALAR

### 3.1.Veri Seti, Kütüphaneler ve Veriyi Görselleştirme

Bu çalışmada 'UCI Machine Learning Repository'de bulunan, 'Fertility Data Seti kullanılmıştır. Bu veri setine UCI veri tabanından ulaşılmıştır (https://archive.ics.uci.edu/ml/datasets/Fertility).

Veri seti içerisinde 100 örnek ve 10 öznitelik bulunmaktadır. 18-36 yaş arasındaki gönüllülerden alınan semen örnekleri, Dünya Sağlık Örgütü (WHO) 2010 kriterlerine göre analiz edilmiştir ve alışkanlıkları ile sağlık durumları hakkında bir form doldurmaları istenmiştir. Veri setinde yer alan öznitelikler ve bunlara karşılık gelen; çeşitli durumlara göre -1 ile 1 arasında bir skor alan değerler Tablo 1'de gösterilmiştir.

Python'da içe aktarma adımlarında, hızlı sayısal dizi hesaplamaları için NumPy v.1.21.5 kütüphanesi, veri analizi için pandas v1.4.4, çizim için matplotlib.pyplot v.3.6.2 ve veri görselleştirme için seaborn v.0.12.1 kullanılmıştır.

Analiz için önce öznitelikleri incelemek amaçlı frekans dağılımı yapılmış sonrasında pandas.plotting kütüphanesi kullanılarak özniteliklerin scatter matrisi oluşturulmuştur.

Ayrıca, çeşitli sınıflandırma algoritmaları, parametreler ve değerlendirme araçları için Python'un ücretsiz makine öğrenmesi kütüphanesinden biri olan Scikit-learn (v.1.2.0) 'den ihtiyaç duyulan kütüphaneler de içe aktarılmıştır.

Support Vector Classifier algoritması için "sklearn.svm" içerisindeki "SVC" fonksiyonu kullanılmıştır. Random Forest için, "sklearn.ensemble" içerisinden "RandomForestClassifier" fonksiyonu; Lojistik Regresyon için "sklearn.linear_model" içerisindeki "LogisticRegression" fonksiyonu; Naive Bayes için "sklearn.naive_bayes" içerisindeki "GaussianNB" fonksiyonu; XGBoost için "xgboost" içerisindeki "XGBClassifier" fonksiyonu; Decision Tree için "sklearn.tree" içerisindeki "DecisionTreeClassifier" fonksiyonu ve son olarak K-en yakın komşu sınıflandırıcı için "sklearn.neighbors" içerisindeki "KNeighborsClassifier" fonksiyonu kullanılmıştır.

**Tablo 1.** Datasetinde yer alan öznitelikler

| Öznitelikler | Değerler |
|---|---|
| **Analizin yapıldığı mevsim** | ✓ kış= -1<br>✓ ilkbahar= -0.33<br>✓ yaz= 0.33<br>✓ sonbahar= 1 |
| **Yaş** | ✓ 18-36 (0, 1) |
| **Çocukluk hastalığı** | ✓ evet= 0<br>✓ hayır= 1 |
| **Kaza ya da travma** | ✓ evet= 0<br>✓ hayır= 1 |
| **Cerrahi müdahale** | ✓ evet= 0<br>✓ hayır= 1 |
| **Yüksek ateş** | ✓ üç aydan kısa bir süre önce= -1<br>✓ üç aydan fazla bir süre önce= 0<br>✓ hayır= 1 |
| **Alkol tüketimi** | ✓ günde birkaç defa, her gün, haftada birkaç defa, haftada bir, neredeyse hiç ya da hiçbir zaman (0, 1) |
| **Sigara tüketimi** | ✓ hiç= -1<br>✓ ara sıra= 0<br>✓ günlük=1 |
| **Günlük oturarak geçen saat** | ✓ 1-16 (0, 1) |
| **Teşhis** | ✓ Normal (N)<br>✓ Farklı (O) |

**3.2.Veri indirme, hazırlama ve işleme aşaması:**

Yapılmış olan bütün analizler Anaconda.Navigator içerisinde bulunan Jupyter Notebook 6.4.12 üzerinden gerçekleştirilmiştir. Veri, UCI üzerinden indirildikten sonra, "path" ve dosya adları tanımlanarak veri kümesi için "Bunch" oluşturulmuştur. Öznitelikler ve etiketler meta datası üzerinden eklenmiştir. Datanın, nokta gösterimi ile erişebilir olması için kütüphanelerden 'Bunch' kullanılarak bunch nesnesi oluşturulmuştur.

Algoritmaları çalıştırmadan önce, eğitim datası ve test datası tanımlanmıştır. "Estimator" ve "predictor" tanımlanarak, dataya uyarlanmıştır. Kullanılacak olan modelleri değerlendirmek için metrikler hesaplanmıştır ve rapor çıkarılmıştır. Son olarak model kaydedilmiştir.

Elde edilen başarım değerlerini arttırmak adına, düzenlenmiş veriler kullanılarak, birkaç farklı sınıflandırıcı (Support Vector Classifier, K-en yakın komşu, Random Forest, Karar Ağacı, Lojistik Regresyon, Naive Bayes ve XGboost) arasında karşılaştırma yapılmıştır.

Modellerin ne kadar doğru çalıştığını anlayabilmek için doğruluk, kesinlik, duyarlılık ve F- ölçütüne bakılmıştır. Yine başarım değerlerini

geliştirmek adına K-katlı validasyon yöntemi kullanılmıştır. Bu çalışma için 12 katlı çapraz doğrulama seçilmiş, test seti kullanılarak 12 kez model değerlendirmesi yapılmıştır. Bu doğrulama yönteminde tüm data seti, rastgele 12 alt kümeye bölünür. Her seferinde datanın bir kısmı test seti, kalanı eğitim datası olarak belirlenir.

## 4.PROBLEM VE KULLANILAN YÖNTEMLER

### 4.1.Kullanılan Algoritmalar

**Support Vector Classifier (SVC);** SVM'ye dayalı olarak geliştirilen destek vektör sınıflandırıcısı (SVC), sayısal örüntü tanıma, yüz algılama, metin kategorizasyonu ve protein katlama tanıma için uygulanmıştır. SVC, hesaplanan girdi verilerinin sınıflandırılması için kullanılmakta olup büyük bir veri seti içeren bir sınıflandırma yöntemidir (Lau & Wu, 2003).

**Random Forest (Rastgele Orman);** denetimli sınıflandırma algoritmalarından biridir. Algoritma, birden fazla karar ağacı üreterek sınıflandırma işlemi esnasında sınıflandırma değerini yükseltmeyi hedefler. Random forest algoritması birbirinden bağımsız olarak çalışan birçok karar ağacının bir araya gelerek aralarından en yüksek puan alan değerin seçilmesi işlemidir (Goos et al., 2012).

**Lojistik Regresyon;** anlamsal olarak birden fazla X'in farklı ikili bir bağımlı değişkenle ilişkisini anlamlandırmak amaçlı kullanılan matematiksel bir modelleme çeşididir (Kleinbaum & Klein, 2010.)Bu regresyon karmaşık fenomenleri anlamak için kullanılmış olan kullanışlı bir teknik olup, (Connelly, L.2020) çıktı değişkeninin olası durumlarını uygun kategoriye bulunup bulunmama ihtimalini hesaplar (Rymarczyk et al., 2019).

**Naive Bayes;** makine öğrenmesi ve veri madenciliğinde kullanılan en işlevsel ve efektif öğrenme algoritmalarından biridir. Sınıflandırmadaki rekabetçi performansı güçlüdür (Zhang, 2004).

**XGBoost,** gözetimli öğrenmenin temel prensiplerinden yola çıkılarak yapılm5ış olup karar ağaçları ve ENSEMBL yöntemlerinin bir araya getirilmesi ile oluşmuş bir yöntemdir. Bu algoritmanın en güzel özelliği yüksek öğrenim gücüne sahip olması, diğer algoritmalara göre daha hızlı olması ve aşırı öğrenmenin önüne geçebilmiş olmasıdır (Brownlee, 2021).

**Decision Tree (Karar ağacı);** öznitelikler ile bağlantılı özellikler hakkında sorular sorarak veri öğelerini sınıflandırır. Her soru bir düğümde bulunmakta ve her dahili düğüm sonucu bir alt düğüm açarak sorunu çözmeyi hedeflemektedir bu sayede sorular bir ağaç şeklinde bir hiyerarşiyi oluşturmaktadır. Karar ağacının avantajı

verilerle ilgili basit soruları anlaşılır bir şekilde birleştirmesidir (Kingsford & Salzberg, 2008).

**K-Nearest Neighbor (KNN, K-en yakın komşu);** etiketlenmemiş her örneği, eğitim setindeki k-en yakın komşuları arasındaki çoğunluk etiketine göre sınıflandırmaktadır. Bu sebepten ötürü performansı, en yakın komşuları belirlemek için kullanılan mesafe metriğine göre değişkenlik göstermektedir (Sun & Huang, 2010).

### 4.2.Kullanılan Metrikler

**Doğruluk (Accuracy)**: bir modelin başarısını ölçmek için çok kullanılan ancak tek başına yeterli olmadığı görülen bir metriktir. Doğru tahminlerin bütün tahminlere bölünmesiyle elde edilmektedir. (Baratloo et al., 2015)

**Kesinlik (Precision):** genellikle bir kişinin bir seferde elde ettiği bir puanın ikinci sefer de tekrarlanma derecesi olarak tanımlanır. Hata matrisindeki nitelikler kullanılarak elde edilen değerdir. Pozitif olarak tahminlenen değerlerin gerçekten kaç adedinin Pozitif olduğunu göstermektedir. Doğru pozitiflik değerinin tahmini pozitiflik değerlerine bölünmesi ile elde edilmektedir (Streiner & Norman, 2006).

**Duyarlılık (Recall):** Doğru pozitifler için doğru tahminlerin yüzdesini gösteren duyarlılık, doğru pozitiflerin, bütün pozitiflere bölünmesi ile elde edilmektedir (Trevethan, 2017).

**F ölçütü;** Recall ve Precision'ın ağırlıklı harmonik ortalamasıdır (Powers, 2019).

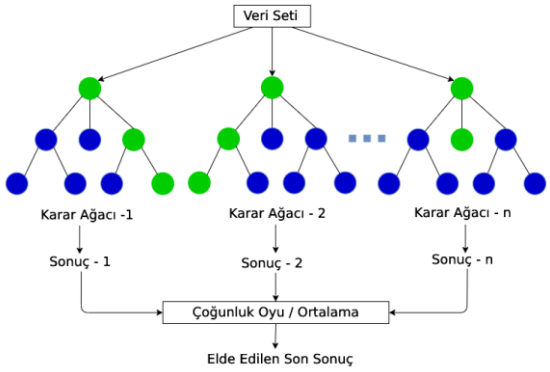
$$F = \frac{2*Recall*precision}{precision+Recall}$$

### 4.3.Önerilen (Geliştirilen/Kullanılan) Yöntem

Bu çalışmada, Support Vector Classifier, K-en yakın komşu, Random Forest, Karar Ağacı, Lojistik Regresyon, Naive Bayes ve XGboost algoritmaları kullanılmak üzere Fertility Data Seti kullanılmış olup içerisinde yer alan öznitelikler incelenmiş, data düzenlenmiş ve eğitim ve test aşamasına hazır hale getirilmiştir. Bahsedilen yedi farklı modeli karşılaştırıp içerisinden en uygun olan algoritmayı seçerek çalışmayı sonlandırmak adına doğruluk, kesinlik, duyarlılık ve F1 ölçütü gibi parametrelerden yararlanılmıştır. Bu sonuçların yer aldığı Tablo 2 aşağıda verilmiştir.

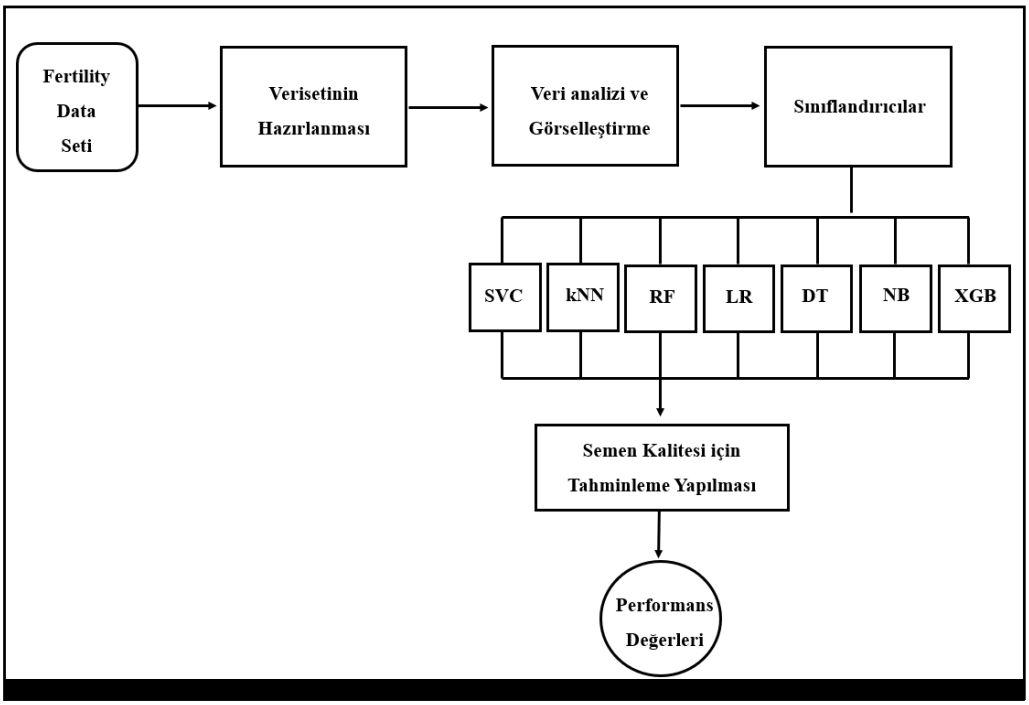
Mevcut yöntemlere kıyasla en yüksek doğruluk ~%88 ile Random Forest sınıflandırma yöntemi ile elde edilmiştir. Doğruluk değeri; modelin başarısını ölçmek için en çok kullanılan metriklerden olmasına rağmen genel bir yorum yapabilmek adına tek başına yeterli olmamaktadır. Bu nedenle duyarlılık ve kesinlik gibi diğer metrikler de

incelenmelidir. Random Forest Algoritması ~%88 duyarlılık ve ~%83 kesinlik değerleri ile bu metrikler için de diğer algoritmaları geride bırakmıştır. Bu doğrultuda duyarlılık ve kesinliğin ağırlıklı harmonik ortalaması olan F1 ölçütü de ~%85 çıkmıştır. Bu çalışmada tüm bu sonuçlar ışığında nihai algoritma olarak Random Forest Sınıflandırıcısı tercih edilmiştir.




**Şekil.1.** Karar Ağacı Şeması (Sarıkay, 2021).

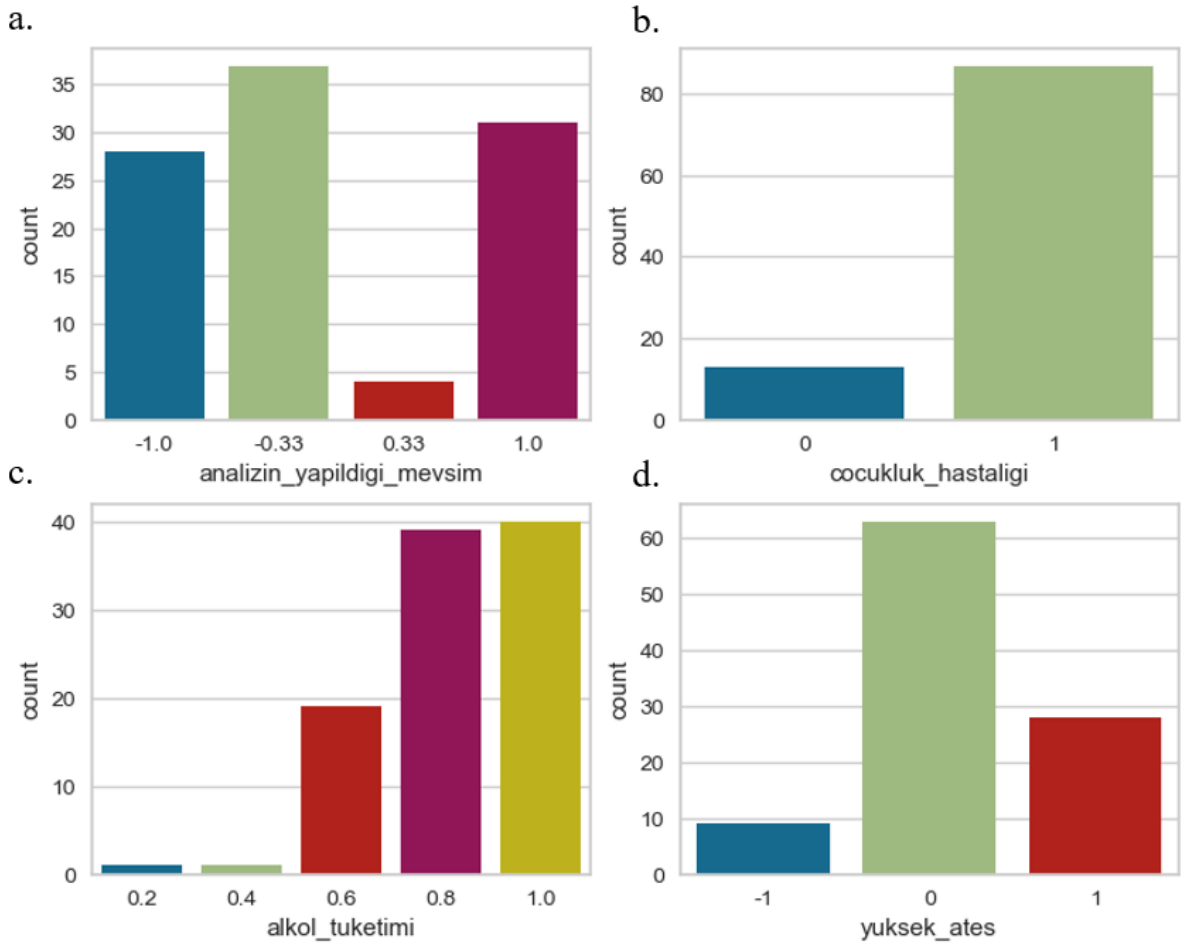
Random forest algoritması, birden fazla karar ağacı üreterek sınıflandırma işlemi esnasında sınıflandırma değerini yükseltmeyi hedefler. Birbirinden bağımsız olarak çalışan birçok karar ağacının bir araya gelerek aralarından en yüksek puan alan değerin seçilmesi işlemidir. Ağaç sayısı ile kesin bir sonuç elde etme oranı doğru orantılıdır. Karar ağaçları algoritması ile arasındaki temel fark, random forest algoritmasında kök düğümü bulma ve düğümleri bölme işleminin rastgele olmasıdır. Avantajlarından biri ise elinde yeterli miktarda ağaç olması halinde overfitting sorununu azaltmasıdır. Şekil 2'de çalışmanın aşamaları, kullanılan yöntemler ve onlar arasından önerilen modelin bulunduğu şema yer almaktadır.




**Şekil.2**. Kullanılan sistem

## 5.BULGULAR VE TARTIŞMA

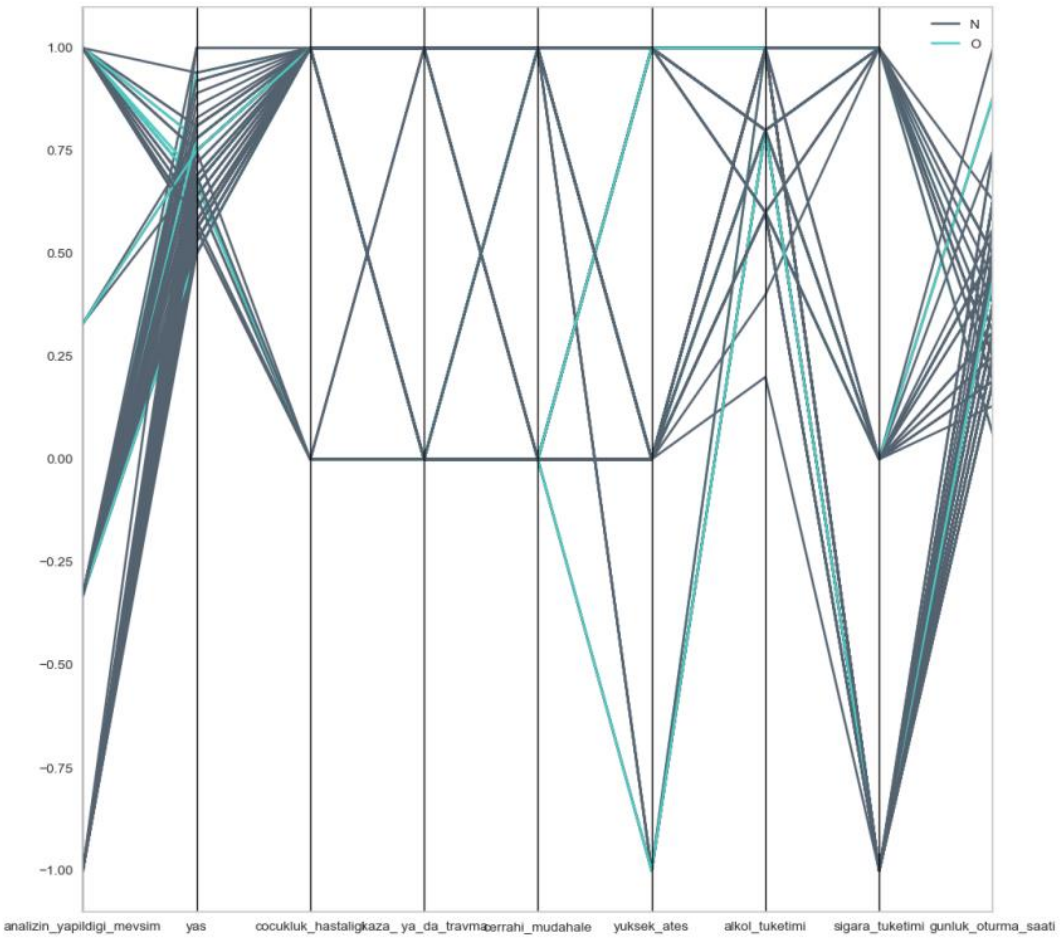
Bu çalışmada, ilk olarak Fertility Data Seti içinde yer alan öznitelikleri daha yakından incelemek için frekans analizi yapılmıştır ve grafikler Şekil 3'de gösterilmiştir. Bu grafikler, çoğu analizin kış, ilkbahar ve sonbahar aylarında yapıldığını, yazın ise çok azının yapıldığını göstermektedir. Çocukluk hastalığı geçirmeyenlerin çoğunlukta olduğu görünmektedir. Ayrıca katılımcıların çoğunun haftada bir alkol tükettiklerini veya neredeyse hiç alkol tüketmedikleri ortaya çıkmıştır. Son olarak son üç aydan önce ateşi çıkmayan katılımcıların sayısı diğerlerine kıyasla daha fazladır.




**Şekil.3.(a)** Analizin yapıldığı mevsimin **(b)** çocukluk hastalığı geçirenlerin **(c)** Alkol tüketenlerin **(d)** yüksek ateşi olanların frekanslar

Bu çalışmada, Fertility Data Seti'nden yararlanarak, çok değişkenli verileri analiz etmek için kullanılan bir görselleştirme tekniği olan "Paralel Koordinatlar Grafiği (PCP)" tercih edilmiştir (Şekil 4). PCP, veri analistlerinin, aralarındaki kalıpları ve ilişkileri arayarak birçok nicel değişkeni birlikte karşılaştırmasına olanak tanır. Bu değişkenler farklı büyüklüklere (farklı ölçekler) ve farklı ölçü birimlerine sahip olduğunda birden fazla sayısal değişkeni aynı anda karşılaştırmak için uygundurlar (Chen et al., 2008).

Görseldeki her satır, veri kümesinden bir örneği ve özniteliklerin her biri için örneğin değerini temsil eder. Renk, teşhis kategorisini temsil eder. Bu bize çeşitli renk kategorilerinin ortak eğilimleri hakkında fikir verir. Grafik incelendiğinde, ara sıra sigara içenlerin (0) sigara içmeyenlere (-1) kıyasla oturarak daha fazla zaman (saat) harcadıkları görünmektedir.




**Şekil.4** Paralel Koordinatlar Grafiği kullanılarak özniteliklerin karşılaştırılması

Aynı zamanda bu çalışmada yedi farklı algoritma karşılaştırılmıştır. Doğruluk, duyarlılık, Kesinlik ve F-ölçütü sınıflandırma yöntemlerinin performansını değerlendirmek için kullanılmıştır. Sonuçlar Tablo 2'de yer almaktadır

Elde edilen sınıflandırma sonuçlarına göre, %84 ile en yüksek F1 değerine sahip makine öğrenmesi algoritması Random Forest sınıflandırıcısı olurken bunu, %82 değeri ile XGBoost, Lojistik Regresyon, SVC ve kNN sınıflandırıcıları takip etmiştir. Naive Bayes'in F1 ölçütü %80 çıkmış, son sırada ise %78 ile Decision Tree yer almıştır. Sonuçların F1 ölçütüne göre yorumlanmasının en temel nedeni, eşit dağılmayan veri kümelerinde hatalı bir model seçimi yapmamaktır.

En iyi sonucu almış olduğumuz Random Forest, tahminin doğruluğuna ulaşmak için tek bir sınıflandırıcı yerine birden fazla sınıflandırıcı kullanarak gelecekteki örnekleri tahmin edebilen kapsamlı bir karar ağacı türüdür (Shaik & Srinivasan, 2019) ve elde ettiği çoklu karar ağaçlarının sonuçlarını, tahmin doğruluğunu geliştirmek için kullanmaktadır. Bu algoritmanın daha yüksek doğruluğa sahip olmasının temeli rastgele oluşundan gelmektedir (Dai et al., 2019). Daha önce de lenf hastalıklarında (Azar et al., 2014), kalp ve damar

**Tablo 2.** Karşılaştırılan algoritmaların performans değerleri

| Algoritmalar | Kesinlik | Duyarlılık | Doğruluk | F1 ölçütü |
|---|---|---|---|---|
| **SVC Classification** | 0.789 | 0.878 | 0.878 | 0.828 |
| **kNN Classification** | 0.784 | 0.878 | 0.878 | 0.825 |
| **Random Forest Classification** | 0.826 | 0.879 | 0.879 | 0.847 |
| **Logistic Regression** | 0.780 | 0.879 | 0.879 | 0.825 |
| **Decision Tree Classifier** | 0.812 | 0.769 | 0.769 | 0.783 |
| **Naive Bayes** | 0.781 | 0.843 | 0.843 | 0.809 |
| **XGBoost Classifier** | 0.818 | 0.849 | 0.849 | 0.829 |

hastalıkları çatısı altında bulunan hastalıkların teşhisinde kullanılan Otomatik Elektrokardiyogram (EKG) yönteminde EKG kalp atışı sinyallerinin sınıflandırılmasında (Alickovic & Subasi, 2016), meme kanseri tahminlemesinde Random Forest algoritması kullanılmış ve başarılı olduğu görülmüştür (Dai et al., 2019). Ayrıca yapılan UCI veri bankasından indirilmiş olan hastalık verilerinin modeller tarafından tahminlemesi sonucu diyabet, koroner kalp hastalığı ve kanser verileri arasında Random Forest modelinin üç hastalık için doğruluk sonucu, Naïve Bayes sınıflandırıcısının doğruluk değerlerinden daha yüksek olduğu görülmüştür (Jackins et al., 2021).

Karşılaştırılan modeller arasındaki XGBoost, Lojistik Regresyon, SVC ve kNN sınıflandırıcıların %82 ile F1 ölçütleri birbirine yakın çıkmıştır. Bu algoritmaların, kesinlik ve duyarlılık sonuçları da birbirine benzerdir. F1 ölçütü kesinlik ve duyarlılığın ağırlıklı harmonik ortalaması olduğu için beklenen bir sonuçtur. Benzer sonuçlar elde edilen algoritmalar arasından ilk olarak Lojistik Regresyon'un sonuçları incelendiğinde doğruluk değerinin %87 olduğu görülmüştür. Lojistik Regresyon Lineer regresyona benzemekte aralarındaki tek fark, değişkenin sonucunun sürekli bir değişken yerine kategorik bir değişken olmasıdır. Sonuçlarımıza benzer bir şekilde, Kalp Hastalığı Veri Kümesi ile yapılan bir çalışmada Lojistik Regresyon'un diğer algoritmalara nazaran en yüksek doğruluğa sahip olduğu bulunmuştur (Kohli Pahulpreet Singh & Arora Shriya, 2018).

Diğer algoritmalara göre daha iyi sonuç aldığımız algoritmalardan biri olan XGBoost (Extreme Gradient Boosting), zayıf öğrenicileri desteklemek için "gradient descent" mimarisini kullanan topluluk ağacı yöntemleridir (Budholiya et al., 2022). Kronik böbrek hastalığı ile ilgili yapılan bir çalışmada XGBoost algoritmasından çıkan skor, şu anda mevcut olan temel modellerden daha iyi olduğu görülmüştür (Ogunleye & Wang, 2018). Her ne kadar XGBoost algoritması yukarıda bahsedilen çalışmalarda diğer algoritmalara göre daha başarılı ve güvenli sonuçlar verse de bu çalışmada XGBoost algoritması Random Forest'a göre daha az başarılı bir tahminleme yapmıştır. Bunun nedeni ağaç tabanlı modellerin herhangi bir sorunun girdi uzayını bölümleme yöntemi kullanması nedeniyle, bu tip algoritmaların tahmin yaparken eğitim verilerinin sınırlarından uzak hedef değerleri büyük ölçüde tahmin edememeleri ve sürekli bir çıktıyı tahmin etmeyi içeren regresyon görevlerinde sınırlamaya sahip olmasıdır. XGBoost karmaşık bir algoritma olmasına rağmen diğer ağaç tabanlı algoritmalar gibi,

tahmin içeren görevler söz konusu olduğunda yetersiz kalmaktadır.

K-en yakın komşu (kNN) algoritması, en basit ve en eski sınıflandırma algoritmalarından biridir. İskemik kalp hastalığına (İKH) sahip, her iki cinsiyetten de karışık 327 kişilik veri seti kullanılarak yapılan araştırmada Karar Ağacı öğrenicileri ve K-en yakın komşular algoritmaları birkaç özellik ile maksimum doğruluk elde etme açısından diğer algoritmalardan üstün olduğu görülmüştür (Kononenko et al., 1999). Bu çalışmada elde edilen sonuçlara göre KNN algoritması sperm kalitesinin tahmini için uygun bir algoritma olduğu ve literatür taraması ile uyumlu olduğu görülmüştür. Bunun nedenin ise veri setinin büyük bir veri kümesi olmaması, çok fazla boyuta sahip olmaması ve aykırı, eksik değerlere sahip olmamasından kaynaklanmaktadır. Bu sayede tahminleme yapılırken maliyet düşük olduğu için performansı yüksek çıkmıştır.

Naïve Bayes (NB) algoritması, Bayes teoremine dayalı bir sınıflandırma tekniğidir. Bu teorem, bir olayın olasılığını, o olayla ilgili koşulların önceden bilinmesine dayalı olarak tanımlayabilir (Uddin et al., 2019). Bu çalışmada 0.80 F-ölçütü ile diğer algoritmalara göre daha az performans gösterdiği görülmüştür. Bu araştırma sonucu kullanılmış olan veri seti ile yapılan çalışmada, Naive Bayes algoritmasının tahminlemesinin bu veri seti için uygun olmadığı bunun nedenlerinden birinin yukarıda belirtilmiş olabileceği gibi veri yetersizliği olabileceği düşünülmüştür.

Karar Ağacı (Decision Tree), en eski ve önde gelen makine öğrenmesi algoritmalarından biridir. Bir karar ağacı, karar mantığını modeller, yani veri öğelerini ağaç benzeri bir yapı içinde sınıflandırmak için sonuçları test eder ve karşılık gelir (Uddin et al., 2019) . Bu çalışmada uygulanmış olan algoritmalar içerisinden en düşük F-ölçütüne, duyarlılığa ve doğruluğa sahip olan Karar Ağacı algoritması bu veri seti için iyi bir tahminleme yapamamıştır. Random Forest; Karar Ağacı uygulamalarından biri olsa da birden fazla karar ağacı üzerinden her bir karar ağacını farklı bir gözlem örneği üzerinde eğiterek farklı modeller üretip sınıflandırma yapmış olduğu için daha derin keşfetme imkânı sunmuş ve veri setini en iyi tahminde bulunan algoritma olmuştur. Aynı şekilde aynı yöntemden çıkmış olan karar ağacı algoritması da bu veri setinin tahminlemesinde yetersiz kalmıştır.

## 6.SONUÇ

Bu çalışmada “Fertility Data Seti” içerisinde yer alan özniteliklerin sperm kalitesi üzerindeki etkisi üzerinden Support Vector Classifier, Karar Ağacı, Lojistik Regresyon, K-en yakın komşu, Naive Bayes ve XGBoost algoritmaları ile tahminleme yapmak amaçlanmıştır.

Sınıflandırma sonucuna göre, 0.84 ile en yüksek F1 değerine sahip makine öğrenmesi algoritması Random Forest olurken bunu, %82 değeri ile XGBoost, Lojistik Regresyon, SVC ve kNN sınıflandırıcılarını takip etmiştir. Naive Bayes %80 ve Karar Ağacı algoritmalarının %78 tahmin değeri ile bu veri seti için diğer algoritmalara göre daha az başarılı tahminde bulunduğu görülmüştür.

Elde edilen tahminler ışığında bu veri seti için bütün değerlerin yüksek olmasından dolayı Random Forest algoritmasının bu çalışmanın sonraki aşamalarında kullanılabileceği görülmüştür, önerilen yöntemin, hastalardaki doğurganlık oranlarını tespit etmede etkili ve doğru sonuçlar verdiğini göstererek, araştırmacılar tarafından önceki çalışmalarda elde edilen sonuçları iyileştirmektedir.

## KAYNAKÇA


- Alickovic, E., & Subasi, A. (2016). Medical Decision Support System for Diagnosis of Heart Arrhythmia using DWT and Random Forests Classifier. *Journal of Medical Systems*, *40*(4), 1–12. https://doi.org/10.1007/s10916-016-0467-8
- Andrade-Rocha, F. T. (2013). *Temporary Impairment of Semen Quality Following Recent Acute Fever*. www.annclinlabsci.org
- Azar, A. T., Elshazly, H. I., Hassanien, A. E., & Elkorany, A. M. (2014). A random forest classifier for lymph diseases. *Computer Methods and Programs in Biomedicine*, *113*(2), 465–473. https://doi.org/10.1016/j.cmpb.2013.11.004
- Babakhanzadeh, E., Nazari, M., Ghasemifar, S., & Khodadadian, A. (2020). Some of the factors involved in male infertility: A prospective review. International Journal of General Medicine, 13, 29–41. https://doi.org/10.2147/IJGM.S241099
- Baratloo, A., Hosseini, M., Negida, A., & Ashal, G. el. (2015). *Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity* (Vol. 3, Issue 2). www.jemerg.com
- Benli, H., HAZNEDAR, B., & KALINLI, A. (2019). Seminal Quality Prediction Using Deep Learning Based on Artificial Intelligence. *Uluslararası Muhendislik Arastirma ve Gelistirme Dergisi*, 350–357. https://doi.org/10.29137/umagd.484786
- Bidgoli, A. A., Komleh, H. E., & Jalaleddin Mousavirad, S. (n.d.). *Seminal Quality Prediction using Optimized Artificial Neural Network with Genetic Algorithm*.
- Brownlee, J. (2021) *XGBoost With Python: Gradient Boosted Trees with XGBoost and scikit-learn*.
- Chen, C., Härdle, W., & Unwin, A. (2008). Handbook of Data Visualization. *Handbook of Data Visualization*. https://doi.org/10.1007/978-3-540-33037-0
- Dai, B., Chen, R. C., Zhu, S. Z., & Zhang, W. W. (2019). Using random forest algorithm for breast cancer diagnosis. *Proceedings - 2018 International Symposium on Computer, Consumer and Control, IS3C 2018*, 449–452. https://doi.org/10.1109/IS3C.2018.00119
- Dash, S. R., & Ray, R. (2020). Predicting seminal quality and its dependence on life style factors through ensemble learning. *International Journal of E-Health and Medical Communications*, *11*(2), 78–95. https://doi.org/10.4018/IJEHMC.2020040105
- Goos, G., Hartmanis, J., Van, J., Board, L. E., Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Kobsa, A., Mattern, F., Zurich, E., Mitchell, J. C., Naor, M., Nierstrasz, O., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., &

Weikum, G. (n.d.). *Information Computing and Applications*.

- Grampurohit Sneha, & Sagarnal Chetan. (2020). *Disease Prediction using Machine Learning Algorithms*.
- Jackins, V., Vimal, S., Kaliappan, M., & Lee, M. Y. (2021). AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. *Journal of Supercomputing*, *77*(5), 5198–5219. https://doi.org/10.1007/s11227-020-03481-x
- Javed Mehedi Shamrat, F. M., Ghosh, P., Sadek, M. H., Kazi, M. A., & Shultana, S. (2020, November 6). Implementation of Machine Learning Algorithms to Detect the Prognosis Rate of Kidney Disease. *2020 IEEE International Conference for Innovation in Technology, INOCON 2020*. https://doi.org/10.1109/INOCON50539.2020.9298026
- K. W., & Wu, Q. H. (2003). Online training of support vector classifier. Pattern Recognition, 36(8), 1913–1920. ttps://doi.org/10.1016/S0031-3203(03)00038-4
- Kingsford, C., & Salzberg, S. L. (2008). *What are decision trees?* http://www.nature.com/naturebiotechnology
- Kleinbaum, D. G., & Klein, M. (n.d.). *Logistic Regression A Self-Learning Text Third Edition*. http://www.springer.com/series/2848
- Kohli Pahulpreet Singh, & Arora Shriya. (2018). *Application of Machine Learning in Disease Prediction* (Kohli Pahulpreet Singh, Ed.).
- Lau, K. W., & Wu, Q. H. (2003). Online training of support vector classifier. Pattern Recognition, 36(8), 1913–1920. https://doi.org/10.1016/S0031-3203(03)00038-4
- Powers, D. M. W. (2019). *What the F-measure doesn't measure… Features, Flaws, Fallacies and Fixes*.
- Princy R.Jane Preetha, Jose P.Subha Hency, Parthasarathy Saravanan, Lakshminarayanan Arun Raj, & Jeganathan Selvaprabu. (2020). *Prediction of Cardiac Disease using Supervised Machine Learning Algorithms*.
- Rymarczyk, T., Kozłowski, E., Kłosowski, G., & Niderla, K. (2019). Logistic regression for machine learning in process tomography. *Sensors (Switzerland)*, *19*(15). https://doi.org/10.3390/s19153400
- Sergerie, M., Mieusset, R., Croute, F., Daudin, M., & Bujan, L. (2007). High risk of temporary alteration of semen parameters after recent acute febrile illness. Fertility and Sterility, 88(4), 970.e1-970.e7. https://doi.org/10.1016/j.fertnstert.2006.12.045
- Shah, D., Patel, S., & Bharti, S. K. (2020). Heart Disease Prediction using Machine Learning Techniques. *SN Computer Science*, *1*(6), 345. https://doi.org/10.1007/s42979-020-00365-y

- Shaik, A. B., & Srinivasan, S. (2019). A brief survey on random forest ensembles in classification model. In *Lecture Notes in Networks and Systems* (Vol. 56, pp. 253–260). Springer. https://doi.org/10.1007/978-981-13-2354-6_27
- Streiner, D. L., & Norman, G. R. (2006). "Precision" and "accuracy": Two terms that are neither. *Journal of Clinical Epidemiology*, *59*(4), 327–330. https://doi.org/10.1016/j.jclinepi.2005.09.005
- Sun, S., & Huang, R. (2010). An adaptive k-nearest neighbor algorithm. *Proceedings - 2010 7th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2010*, *1*, 91–94. https://doi.org/10.1109/FSKD.2010.5569740
- Swaminathan, S., Qirko, K., Smith, T., Corcoran, E., Wysham, N. G., Bazaz, G., Kappel, G., & Gerber, A. N. (2017). A machine learning approach to triaging patients with chronic obstructive pulmonary disease. *PLoS ONE*, *12*(11). https://doi.org/10.1371/journal.pone.0188532
- Trevethan, R. (2017). Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice. *Frontiers in Public Health*, *5*. https://doi.org/10.3389/fpubh.2017.00307
- Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, *19*(1). https://doi.org/10.1186/s12911-019-1004-8
- Virtanen, H. E., Jørgensen, N., & Toppari, J. (2017). Semen quality in the 21 st century. In Nature Reviews Urology (Vol. 14, Issue 2, pp. 120–130). Nature Publishing Group. https://doi.org/10.1038/nrurol.2016.261
- Wang, R., Zhou, H., Zhang, Z., Dai, R., Geng, D., & Liu, R. (2013). The impact of semen quality, occupational exposure to environmental factors and lifestyle on recurrent pregnancy loss. *Journal of Assisted Reproduction and Genetics*, *30*(11), 1513–1518. https://doi.org/10.1007/s10815-013-0091-1
- Wiwanitkit, V. (2010). Influenza, swine flu, sperm quality and infertility: A story. In J Hum Reprod Sci (Vol. 3, Issue 2, pp. 116–117). https://doi.org/10.4103/0974-1208.69339
- Wu, C. C., Yeh, W. C., Hsu, W. D., Islam, M. M., Nguyen, P. A. (Alex), Poly, T. N., Wang, Y. C., Yang, H. C., & (Jack) Li, Y. C. (2019). Prediction of fatty liver disease using machine learning algorithms. *Computer Methods and Programs in Biomedicine*, *170*, 23–29. https://doi.org/10.1016/j.cmpb.2018.12.032
- Yogev, L., Kleiman, S., Shabtai, E., Botchan, A., Gamzu, R., Paz, G., Yavetz, H., & Hauser, R. (2004).

Seasonal variations in pre- and post-thaw donor sperm quality. *Human Reproduction*, *19*(4), 880–885. https://doi.org/10.1093/humrep/deh165

- You, J. B., McCallum, C., Wang, Y., Riordon, J., Nosrati, R., & Sinton, D. (2021). Machine learning for sperm selection. In *Nature Reviews Urology* (Vol. 18, Issue 7, pp. 387–403). Nature Research. https://doi.org/10.1038/s41585-021-00465-1
- Zhang, H. (n.d.). The Optimality of Naive Bayes. www.aaai.org