YILDIZ TEKNİK ÜNİVERSİTESİ
BİLGİSAYAR MÜHENDİSLİĞİ

VERİ MADENCİLİĞİ PROJESİ

BERKAY KOÇ - 18011047

SERKAN YILDIRIM - 17011078

2021

İÇİNDEKİLER

1. GİRİŞ

Veri madenciliği dersi projesi için yapılan bu ödevde amaç, suların içerisindeki aşağıda, giriş kısmının devamında daha detaylıca değinilecek olan farklı dokuz bileşenin çeşitli veri madenciliği metotları ile değerlendirilerek suyun içilip içilemeyeceği ile ilgili yüksek oranlı tahminler çıkartmaya çalışmaktır. İçilebilir su, içmek veya yemek hazırlamak için güvenilir, içildiğinde insan sağlığına herhangi bir yan etkisi olmayan sudur. ABD Ulusal Bilimler, Mühendislik ve Tıp Akademileri, günlük yeterli sıvı alımını, ılıman bir iklimde yaşayan ortalama, sağlıklı yetişkin bir erkek için yaklaşık 3,7 litre, bir kadın için ise 2,7 litre olarak belirlemiştir. Bunun yanı sıra sıcak iklimlerde kaybedilen su miktarının artması dolayısıyla alınması gereken su miktarının da gerekliliği aşikârdır. Bununla beraber dünya üzerinde yaklaşık 3 milyar insanın yeterli miktarda sağlıklı içme suyuna erişmekte sıkıntı çektiği bilinmektedir.[1] Dünya üzerinde yaklaşık 2 milyar insanın ise dışkıyla temas etmiş bir içme suyu kullandığı bilinmektedir.[1] Kullanılan kirli suların ishal, kolera, dizanteri, tifo ve çocuk felci başta olmak üzere bir çok hastalığı bulaştırabildiği bilinmektedir.[2] Bu bilgiler ışığında içilebilir suyun insanlık için ne kadar kıymetli olduğu bellidir. İçilebilir ve sağlıklı suyun belirlenmesi için ise yapılan araştırmalar sonucunda elde edilmiş verilerin ölçüldüğü içilebilirlik testleri kullanılmaktadır.

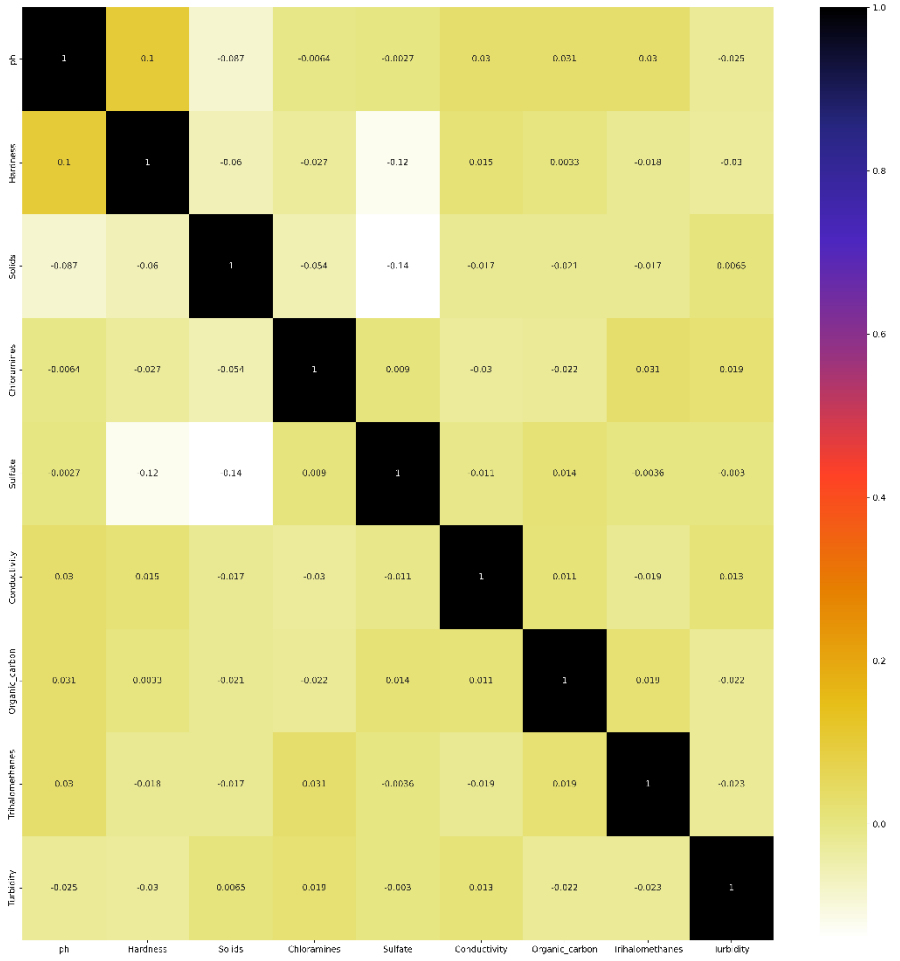
Projede kullanılan veri seti https://www.kaggle.com/artimule/drinking-water-probability adresinden erişilebilen, herkese açık olarak paylaşılmış olan veri setidir. Veri setinde 10 adet özellik bulunmaktadır. Bu özellikler şöyledir;

- ph: suyun ph değeridir.
- Hardness: Suyun kalsiyum ve magnezyum tuzlarından kaynaklanan sertliğini belirtir.
- Solids: Potasyum, sodyum, kalsiyum, magnezyum gibi minerallerin çözünebilme derecesidir.
- Chloramines: Sudaki klorin ve kloramin oranlarını ifade eden sayısal değerdir.
- Sulfate: Topraklarda, kayalarda ve minerallerde bulunan sülfatların suda bulunma oranıdır.
- Conductivity: Suyun iletkenliğini ifade eden sayısal değerdir.
- Organic carbon: Saf sudaki organik bileşiklerdeki toplam karbon miktarıdır.
- Trihalomethanes: Trihalometanlar, klor ile işlenmiş suda bulunabilen kimyasallardır. Bu değer de sudaki trihalometan miktarını belirtir.
- Turbidity: Suyun bulanıklık değerini ifade eder. 0 ile 1000 arasında ölçüm yapabilir fakat DSÖ 5 üzerindeki turbidity değerlerini sağlıksız kabul eder.
- Potability: Suyun yukarıdaki özellikler analiz edilerek elde edilen içilebilirliğidir. 0 içilemez, 1 içilebilir demektir.

Veri setinin içerisinde 1265'i eksik veriler barındırmak üzere toplamda 3276 veri objesi bulunmaktadır. Bu veri objeleri proje boyunca daha iyi tahmin sonucu elde etmek adına bazen eksik değerler arındırılarak kullanılmış, bazen ise çeşitli eksik veri doldurma metotlarına başvurulmuştur. Metot kısmında bu iki farklı yaklaşım arasındaki ayrım ve tahmin başarı oranları detaylıca incelenecektir.

Bu bağlamda içilebilirlik değerlerinin tahmini için kullanılacak olan metotlar daha önce sunulmuş ara raporda da belirtildiği gibi Karar Ağaçları, Naive Bayes Metodu ve Rastgele Orman algoritmaları olacaktır.

Veriyi tanımak ve veri ile ilgili bilgi edinmek için korelasyon matrisi ve kutu grafiği çizdirilmiştir.



Resim 1. Öznitelikler arası korelasyon matrisi

Resim 1.de, potability dışındaki verilerin birbirleriyle olan korealasyon matrisi çıkartılmıştır. Burada kolayca görülebileceği üzere verilerin kendi aralarında pozitif veya negatif diyebileceğimiz bir benzerlik yoktur. Bu nedenle verideki her özniteliğin farklı anlamlara geldiği, birinin değerinin diğerinin değerine bir etkisi olmadığı söylenebilir. Bu yüzden verideki öznitelikler için herhangi ekstra bir işlem yapma gereksinimi duyulmamıştır. Proje için kullanılmış veri seti ile ilgili göze çarpan bir başka önemli nokta ise verinin dengesiz oluşudur. Potability değerinin 1 olduğu 1198 değer varken 0 olduğu 1930 değer

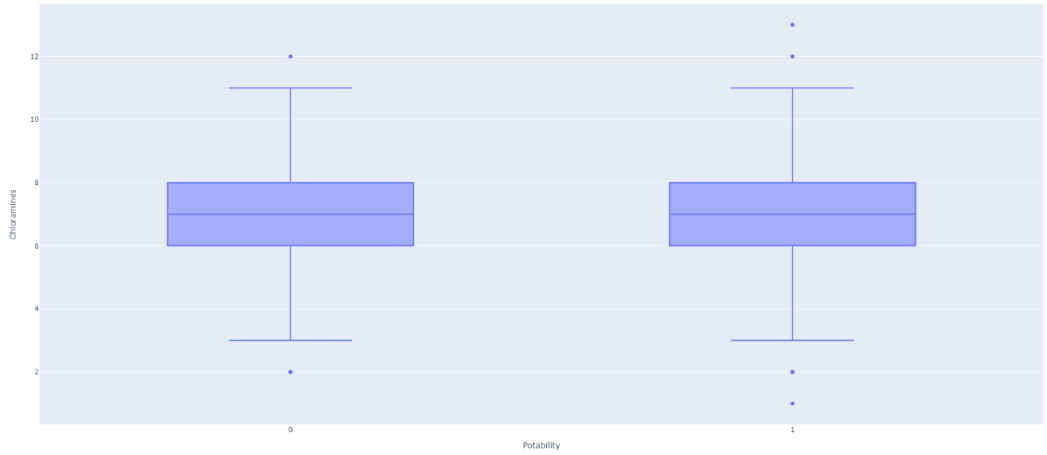
vardır. Temizlenmiş veri setinde ise bu sayılar 1 olduğu değer için 1167 olurken 0 olan değerler için 766'dır. Bu da çalışılan veri setini oldukça dengesiz hale getirmektedir ve tahmin sonuçlarına muhakkak etkisi olacaktır.

Öznitelikleri temsil eden kutu grafikleri ise şöyledir;

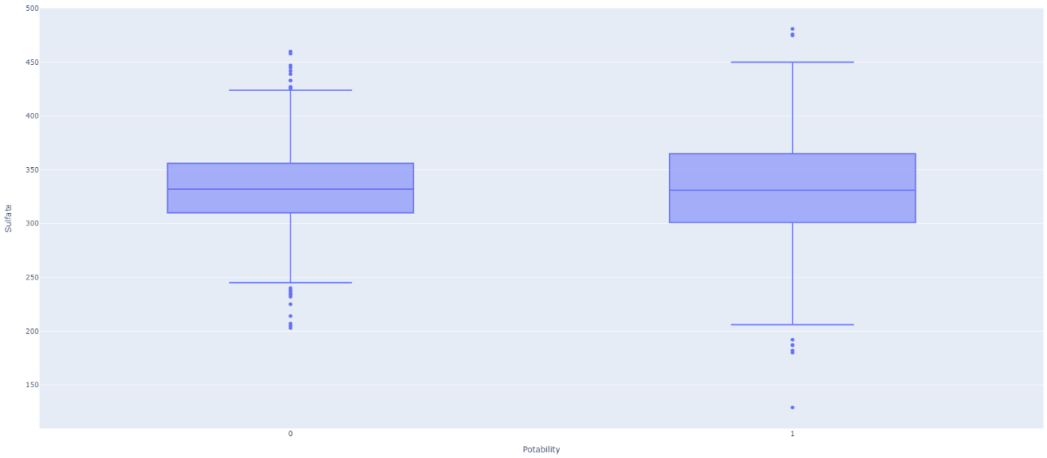
Resim 2. Hardness özniteliği için kutu grafiği

Resim 3. Solids özniteliği için kutu grafiği

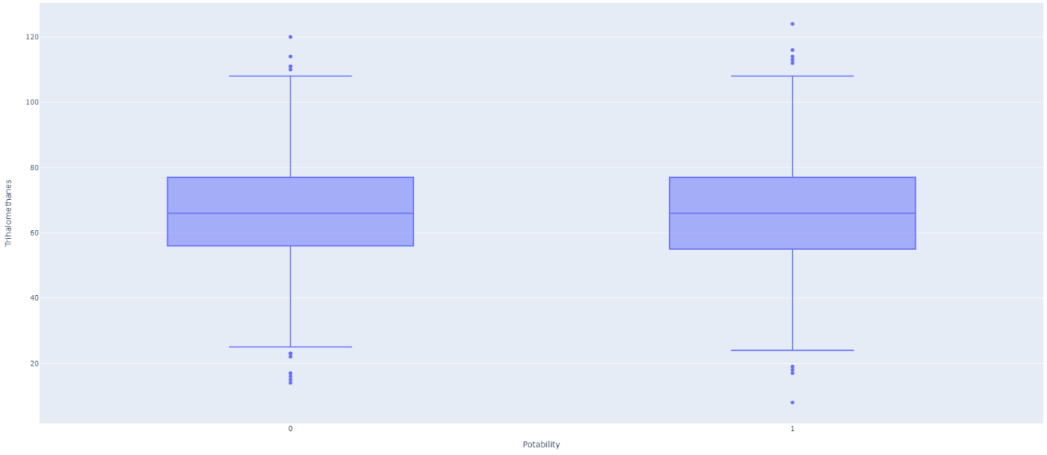
Resim 4. ph özniteliği için kutu grafiği



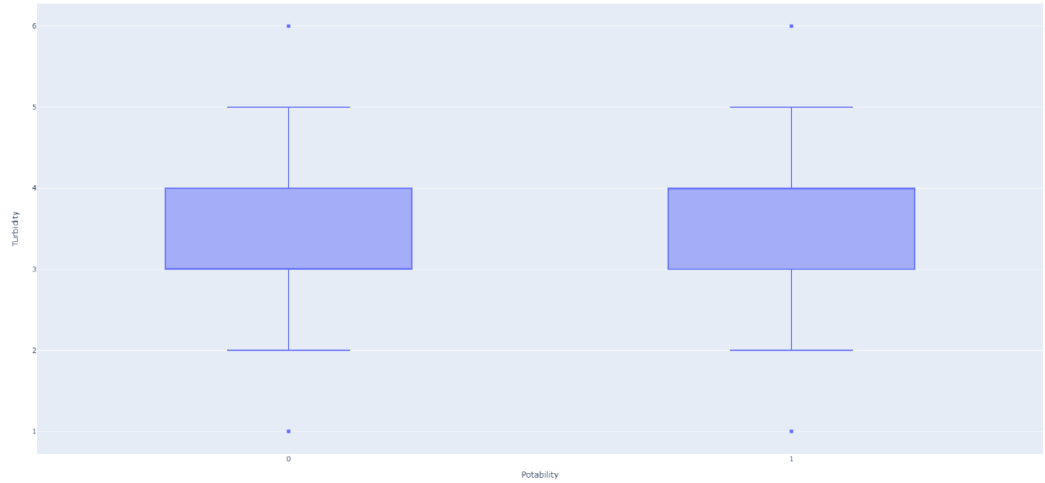
Resim 5. Chloramines özniteliği için kutu grafiği



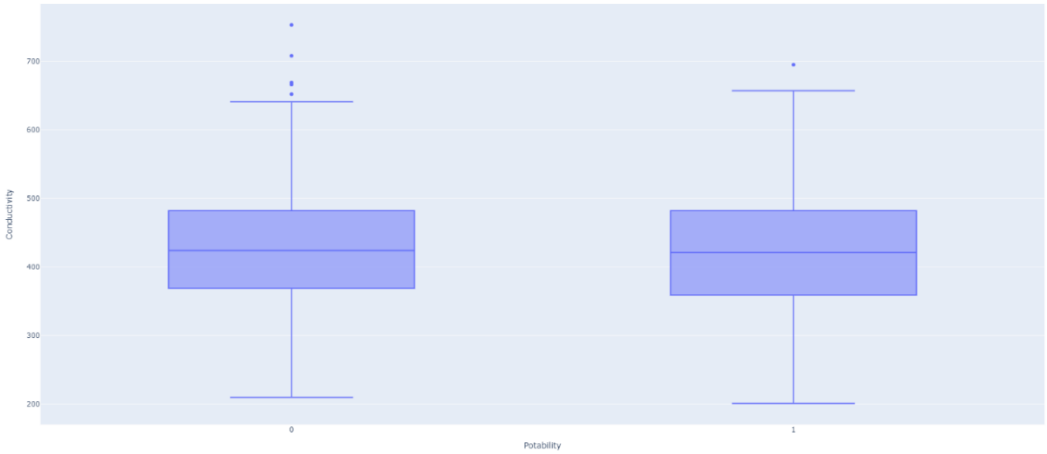
Resim 6. Sulfate özniteliği için kutu grafiği



Resim 7. Trihalomethanes özniteliği için kutu grafiği



Resim 8. Turbidity özniteliği için kutu grafiği



Resim 9. Conductivity özniteliği için kutu grafiği

Resim 10. Organic Carbon özniteliği için kutu grafiği

Kutu grafiklerine bakıldığında medyan değerler, çeyrekler açıklığı değerleri ve Kutu grafikleri yorumlandığında ise, Chloramines, Conductivity ve Turbidity öznitelikleri için çok fazla

outlier değer olmaması fakat diğer özniteliklerde çok fazla outlier değer olması göze çarpmaktadır.

Korelasyon matrisi ve kutu grafikleri verinin tanınmasını sağlamış, verinin uygun hale getirilmesi için yapılması gerekenleri belirgin hale getirmiştir.

[1]. https://www.who.int/en/news-room/fact-sheets/detail/drinking-water
[2]. https://en.wikipedia.org/wiki/Drinking_water

2. ÖNİŞLEME

Verinin giriş kısmında detaylıca değerlendirilmiş olan özellikleri neticesinde veride özellikle bazı özniteliklerde fazlaca bulunan aykırı değerlerin veri setinden arındırılması işlemine başlanmıştır. Bu işlem için ise Z-score normalizasyonu yöntemi uygulanmış, anormal değerler veri setinden atılmıştır.

Bu işlem için öncelikle tüm öznitelikler için z-score normalizasyonu değerleri hesaplanmış, elde edilen verilerden -3 ile +3 arasında olmayan, yani aykırı kabul edilen değerler atılmıştır.

Z-score normalizasyonu ile veri seti temizleme işlemi temizlenmiş veri seti ve ortalamaları alınarak doldurulmuş veri seti için ayrı ayrı yapılmıştır. Her iki durum için de aykırı değerler tahmin başarısını etkileyeceğinden bu yönteme başvurulmuştur. Boş değerlerin temizlendiği veri setinde oldukça düşük miktarda aykırı değere rastlanmış, veri seti 2011 veriden 1933 veriye düşmüş, ortalama alınarak doldurulanda ise veri seti 3276 veriden 3128 veriye düşmüştür.

3. METOT

Kullanılacak olan metotlar Karar Ağaçları, Naive Bayes ve Rastgele Orman algoritmasıdır. Bu metotlarla yaptığımız işlemlere değinmeden önce projede kullanılmış olan metotlar ile ilgili genel bilgiler verilecektir.

3.1 KARAR AĞAÇLARI

Karar ağaçları, makine öğrenmesi ve veri madenciliği alanlarında kullanılan, hem sınıflandırma problemleri hem de regresyon problemleri için kullanılabilen bir tür parametrik olmayan modeldir. Modeldeki esas amaç özniteliklerin değerlerine Düğüm ve

dalları kullanarak düğümlerin öznitelikler, dalların özniteliklerin olabileceği değerler, yaprakların ise tahmin sonuçları olduğu bir ağaç yapısı tasarlamaktır. Karar ağaçları yapısı oluşturulurken bir dizi öznitelik arasından hangisinin düğüm olacağı, hangi değerlere göre dallanmalar olacağı ve bu dallanmaların hangi düğümler tarafından takip edileceği ve en nihayetinde tahminin ne olacağını belirlemek için en bilinen üç yöntem vardır. Bu yöntemler entropi, bilgi kazancı ve gini indeksidir. Bu üç metodun özniteliklere uygulanmasıyla en verimli yolun bulunması hedeflenir. Özniteliklere bir dizi işlem uygulanarak elde edilen sayısal sonuçlar değerlendirilerek seçimler yapılır.

3.1.1 ENTROPİ

Ağacı düğümlere bölmek için dikkat edilen özelliklerden en önemlisi düğümün homojen olmasıdır. Değerlerin homojenliğinin hesaplanması için ise entropi yöntemi kullanılır. Eğer hesapladığımız özniteliğimiz tamamen homojense entropi sıfırdır ve numune eşit olarak bölünmüşse yani heterojen yapıdaysa entropisi birdir. Entropi hesabı için

$$E(S) = \sum_{i=1}^{c} -p_i \, log_2 pi$$

formülü kullanılır.

3.1.2 BİLGİ KAZANIMI

Bilgi kazanımı, bir özniteliğin bir sınıf hakkında ne kadar bilgi sağladığını ifade eden ölçü olarak tanımlanabilir[3]. Bir karar ağacın düğümlerine yerleşecek olan özniteliklerin hangi sırada olması gerektiğini belirlemeye yardımcı olur. Bir karar ağacındaki düğümlerin ne kadar iyi bölündüğünü belirlemek için bilgi kazancı hesabı kullanılabilir. Bilgi kazancı için

$$Info(D) = -\sum_{i=1}^{m} p_i \, log_2(pi)$$

formülü kullanılır.

3.1.3 GINI İNDEKSİ

Gini safsızlığı olarak da bilinen Gini indeksi, rastgele seçildiğinde yanlış sınıflandırılan belirli bir özelliğin olasılık miktarını hesaplar.[3] Her bir sınıfın olasılıklarının karelerinin toplamının birden çıkarılması ile elde edilir. Hesaplanması için

$$GINI(t) = 1 - \sum_{j} [p(j|t)]^2$$

formülü kullanılır.

3.2 NAIVE BAYES

Naive Bayes, makine öğrenmesinde ve veri madenciliğinde sıklıkla kullanılan bir sınıflandırma algoritmasıdır.[5] Algoritma, ismini de aldığı Thomas Bayes'in sunduğu Bayes Teoremine dayandırılmaktadır. Dolayısıyla Naive Bayes sınıflandırmasına değinmeden evvel Bayes Teoremi'ni ve ne için kullanıldığını anlamak gerekir.

Bayes Teoremi koşullu olasılıkların hesaplanmasını sağlayan bir yöntemdir. Başka bir durumun gerçekleştiği göz önüne alındığında bir diğer olayın olma olasılığıdır. Koşullu olasılığı kullanarak, bir önceki olayın bilgisi göz önüne alındığında, bir olayın gerçekleşme olasılığını bulabilir. Buradaki önemli nokta bu iki olayın birbirinden bağımsız olaylar olması gerekliliğidir. B olayı olduğunda A olayının olma olasılığını ifade etmek için P(A|B) gösterimi kullanılır ve olasılığı

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Formülüyle hesaplanır.

3.3 RASTGELE ORMAN

Rastgele Orman algoritması, topluluk metotlarından biridir. Bu algoritma özünde birden fazla kez denemeyle bu denemeler sonucunda elde edilmiş en iyi modeli seçme mantığına dayanır. Rastgele Orman metodu bireysel tahminlerden ziyade toplu tahminlerin iyi çıkmasını hedefler. Oluşturulan bazı ağaç modelleri az başarılı sonuç verirken, daha sonrasında elde edilecek olan sonuçların daha başarılı sonuç vermesini sağlarlar. Bir grup olarak çalışan çok sayıda modelin olması bu sebeple tercih edilmektedir. Bunun yanı sıra Rastgele Orman algoritması tüm bunları yapmaya çalıştığı için fazlaca işlem yapmaktadır, bu sebeple çalışma süresi fazlaca uzayacaktır.

Bir Rastgele Orman algoritmasının daha iyi çalışabilmesi için oluşturulan ağaçların ürettikleri tahminlerin birbiriyle düşük korelasyona sahip olması istenir. Bunun sebebi oluşturulan algoritmanın daha farklı ağaçlara bakarak daha iyi yorumlamalar yapmasını sağlamaktır.

Rastgele Orman algoritması Karar Ağaçları metodunun getirdiği aşırı uyum gösterme problemine de çözüm bulur. Oluşturulan modelde fazlaca sayıda oluşturulan ağaç tarafından elde edilen bilgi bulunduğundan aşırı uyum gösterme noktasında bir sorun yaşamayacaktır.

4. UYGULAMALAR

Bir sınıflandırma modeli oluşturduktan sonra, o model tarafından yapılan tahminlerin ne kadar iyi olduğunu değerlendirmeniz gerekir.[4] Bu projede uygulanan üç metot (Karar Ağaçları, Naive Bayes, Rastgele Orman) hem ilk başta bütün kayıp verilerin bulunduğu satırların silinerek oluşturulan veri kümesinde, hem de kayıp veriler o özelliğin ortalamaları ile doldurulmuş olan veri kümesinde uygulanmıştır. Böylelikle hem bu iki veri kümeleri

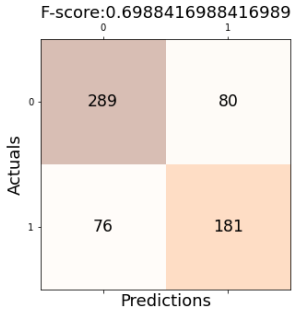
arasında oluşan fark hem de metotlar uygulandıktan sonra oluşan fark gözlemlenebilmektedir.

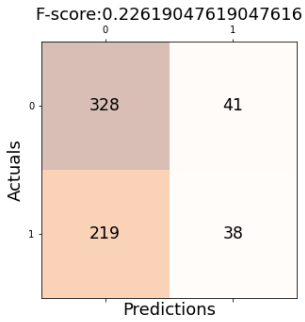
4.1 F-score

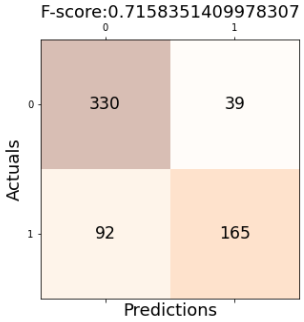
F-score metriği, uygulanan modelden elde edilen sonuçlarda Precision ve Recall değerlerinin harmonic ortalamaları hesaplanarak oluşturulur. F1-Score = 2* Precision Score * Recall Score/ (Precision Score + Recall Score) formülünde kullanılması ile elde edilir. Modelden elde edilen çıktılardaki True Positive, False Positive, True Negative, False Negative değerlerinin karşılaştırılmasını sağlar. Doğruluk ve F-score değerleri farklı kavramlardır.Analiz edilecek veri kümesine bağlı olarak ikisinden biri kullanılabilir. F-score, False Negative ve False Positive değerleri model için çok önemli olduğu durumlarda kullanılır. Çoğu gerçek sınıflandırma problemleri dengesiz sınıfları bulunan veri setleri barındırdığı için F-score uygulamak daha iyi bir yöntemdir. Modelimiz False Negative değerini azaltmak amaçlı oluşturulduysa F-score kullanılması gerekmektedir.

4.2 Karmaşıklık Matrisi

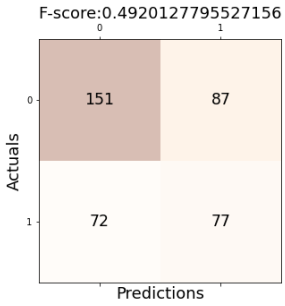
Karmaşıklık matrisi bir tahminleyici tarafından yapılan tahminlerin tabloya aktarılmış özet halidir.[7] Karmaşıklık matrisi modelden elde edilen sonuçlardaki True Positive, False Positive, True Negative, False Negative değerlerinin bir matris üzerinde gösterilmesi ile elde edilir. Modelin nasıl bir doğrulukta çalıştığının gözlemlenmesi için kullanılır. Karar Ağaçları, Naive Bayes, Rastgele Orman metotları kullanılarak kayıp veriler ortalama değerler ile doldurularak oluşturulan veri kümesi ile oluşturulan modellerin F-score ve Karmaşıklık matrisleri sırasıyla:

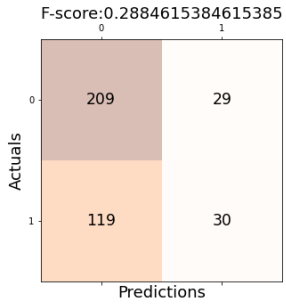


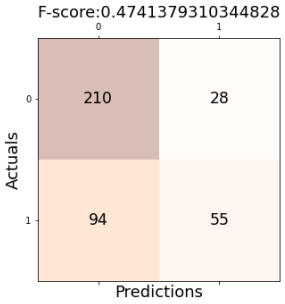




Karar Ağaçları, Naive Bayes, Rastgele Orman metotları kullanılarak kayıp veriler silinerek oluşturulan veri kümesi ile oluşturulan modellerin F-score ve Karmaşıklık matrisleri sırasıyla:





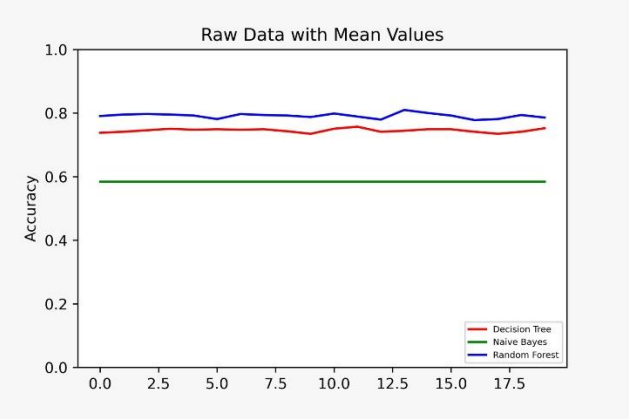


4.3 Doğruluk

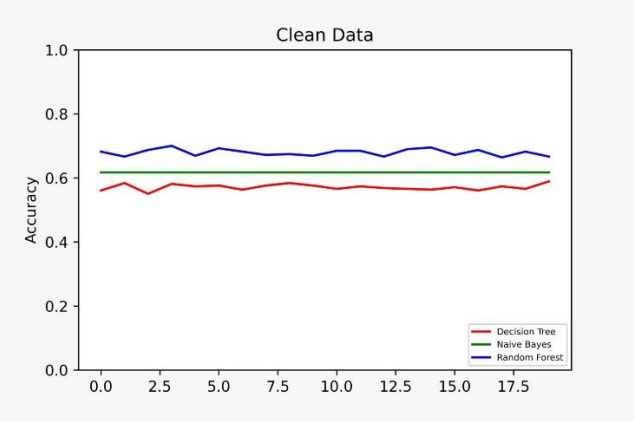
Doğruluk en basit şekilde modelin nasıl çalıştığını belirten metriktir.

$$Acc(M) = \text{True Positive} + \text{True Negative} / \text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative}$$

formülü ile bulunur. Sınıflar eşit derecede önemli olduğu durumlarda kullanılabilir. Projemizde oluşturduğumuz modellerin doğrulukları Karar Ağaçları, Naive Bayes, Rastgele Orman metotları kullanılarak kayıp veriler ortalama değerler ile doldurularak oluşturulan veri kümesi ile oluşturulan modellerin doğrulukları şöyledir;



Karar Ağaçları, Naive Bayes, Rastgele Orman metotları kullanılarak ve kayıp veriler silinerek oluşturulan veri kümesi ile oluşturulan modellerin doğrulukları ise şöyledir;



5. TARTIŞMA

Bu kısımda metotların ayrı ayrı başarı tahminlerine bakılacak, bu değerlendirmeler arasında diğer metotlarla karşılaştırmaları yapılacaktır.

5.1 KARAR AĞAÇLARI

Karar ağaçları yöntemi boş değerlerin bulunduğu veri objelerinin tamamen silindiği durumda yaklaşık 0,6'lık bir tahmin başarısı yakalarken verinin ortalama değerler kullanılarak doldurulduğu durumda tahmin başarısı 0,8 civarlarına yükselmiştir. Bunun sebebini verinin dengesiz olmasına ve temizlenmiş veri seti kullanıldığında veri sayısının azalmasına bağlayabiliriz.

5.2 NAIVE BAYES

Naive Bayes metodunun yazılı metinlerden oluşan veri setlerinde ve çok sayıda sınıf tahmini gerektiğinde iyi çalışması gerektiği bilinmektedir.[6] Veri setimizde bu gibi özellikler olmadığından Naive Bayes tahminleyicisi 0,6 tahmin başarısından öteye geçememiştir. Her iki veri setinde de 0,6 değerini bulması konusunda, her iki veri setinin de tahmin edilecek olan potability özniteliğinin değer dağılımının yaklaşık olmasıyla açıklayabiliriz.

5.3 RASTGELE ORMAN

Rastgele orman metodunun her iki veri setinde de en iyi dereceleri verdiği uygulamalar bölümündeki doğruluk alt başlığındaki grafiklere bakılarak görülebilir. Temiz verilerden oluşan veri setinde yaklaşık 0,6 tahmin başarısı verirken doldurulmuş verilerle oluşturulmuş veri setinde 0,8'lik tahmin başarısı yakalamıştır. Bunun sebebi olarak rastgele orman algoritmasının birden fazla kez karar ağacı oluşturarak sonuçta çok daha iyi bir karar ağacı elde etmeye dayanan bir algoritma olmasını söyleyebiliriz.

6. İŞ DAĞILIMI

Projenin tüm aşamaları ekip üyelerinin ortak katkısıyla yapılmıştır.

7.SONUÇLAR

Süreç boyunca genel hatlarıyla bir veri setinin nasıl önişleme adımlarına tabii tutulması gerektiği, verinin nasıl analiz edileceği, hangi aşamada hangi metodun kullanılması gerektiğine nasıl karar verilmesi gerektiği anlaşılmıştır. Uygulanan metotların hazırlanan iki

farklı tipteki aynı veri seti üzerinde nasıl sonuçlar verdiği tartışma kısmında da incelendiği gibi yorumlanmış, bu üç metotla ilgili derin bilgi sahibi olunmuştur.

REFERANSLAR

[1]. https://www.who.int/en/news-room/fact-sheets/detail/drinking-water
[2]. https://en.wikipedia.org/wiki/Drinking_water
[3]. https://medium.com/analytics-steps/understanding-the-gini-index-and-information-gain-in-decision-trees-ab4720518ba8
[4]. https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2

[5]. https://medium.datadriveninvestor.com/a-gentle-introduction-to-naive-bayes-classifier-9d7c4256c999

[6]. https://www.quora.com/When-and-why-is-a-naive-Bayes-classifier-a-better-worse-choice-than-a-random-forest-classifier/answer/Eren-Golge?ch=10&oid=2785164&share=430a4a72&target_type=answer

[7]. https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5