Dupli Checker

PLAGIARISM SCAN REPORT

0%
Plagiarised

100%
Unique

Date	2023-05-18
Words	927
Characters	8052

Content Checked For Plagiarism

GİRİŞ

İnsanlık tarihine bakıldığında, ilk başta hayatta kalmak, yemek bulabilmek gibi temel ihtiyaçların karşılanması adına insanların, yaşanılan çevreyi detaylı bir şekilde incelenerek o çevreye uyum sağlamaya çalışma amacı fark edilebilir. Bunun yapılabilmesi için insanlar etrafı inceleyerek veri sağladılar ve bu verileri görerek, izleyerek ve deneyerek tecrübelendiler. Bu şekilde, eldeki veriden hareketle ilgili konuda bilgi sağlayıp hayatlarına o bilgiyi entegre ederek yaşadıkları görülmektedir. Günümüz dünyasına geldiğimiz zaman, veriden bilgiye giden sürecin felsefesi aynı iken ilgili verileri elde etme, analiz etme ve bilgiye dönüştürme şekillerinde farklılıkların söz konusu olduğu görülmektedir. Bu durum, başta bilgi elde etmenin ve bunun hemen öncesinde veriyi işleyebilmenin ne kadar büyük bir öneme sahip olduğunu göstermektedir.

Veriyi inceleme ve işleme yani analize hazır hale getirme sürecinde, veride yer alan birtakım problemler, analize engel olacak veya analizin sağlıklı yapılabilmesine engel olabilecek durumlar söz konusu olabilmektedir. Bunlardan birisi ve en geneli verilerde eksik değerlerin bulunmasıdır. İlgili veri setinin analiz edilebilmesi için bahsi geçen eksikliklerin üzerine eğilmek ve mümkün sınırlar çerçevesinde bu problemin ortadan kaldırılması gerekmektedir. Pratikte, en genel ve ilk akla gelen yaklaşımın eksik değerlerin olduğu verileri silmek diğer bir ifade ile yok saymak olduğu görülmüştür. Bu durum bazı durumlarda problem teşkil etmez iken aksi durumlarda, genel olarak düşünüldüğünde veri setinin yapısının, doğasının dolayısıyla karakteristik özelliklerinin farklılaşması, bozulmasını anlamına gelebilir. Bu nedenle, veri setindeki eksik değer oranı bu konuda bir ön fikir vermektedir. Eğer ki bu oranın yüksek sayılabilecek bir oranda iken eksik değerlerin olduğu verilerinin silinmesi gibi bir yol izlenirse elde edilen veri seti ile ilk baştaki veri seti istatistiksel olarak karşılaştırıldığında neredeyse bambaşka iki farklı veri seti ile karşı karşıya olunduğu bir durum oluşabilecektir.

Bu noktada veri setinin genel özelliklerini taşıyan, veri setinin yapısını bozmayan bu anlamda uygun değerler ile eksik olan kısımlar tamamlanmalıdır. Belirtilen bu durumda, eksik verileri tamamlama çerçevesi altında birçok yöntem barındırmaktadır. Bu yöntemler, verinin özelliklerine göre farklılık göstermektedir. Örneğin; en çok tercih edilen temel yöntemlerden ortalama atama, medyan atama veya eksik değerlerin bulunduğu veri kayıtlarını silmenin yanı sıra makine öğrenmesi algoritmaları da eksik değerlerin tamamlanmasında kullanılmaktadır. Bu yöntemleri kullanarak verinin yapısına uygun bir şekilde eksik verilerin tamamlanması yapılmaktadır.

1. Eksik Veri ve Türleri

Yapılan araştırmalara bakıldığın eksik verilerin bulunması oldukça sık karşılaşılan bir durumdur. Var olan bu eksiklikler elbette birçok sebepten kaynaklı olabilmektedir. Bu noktada, eksik verilerin eksik olma sebeplerine dair bilgi edinip açıklama yapmak için istatistiksel ve matematiksel bir bakış açısıyla eksik veri türleri ortaya koyulmuştur. Bu türler üç başlık altında MCAR(Missing Completely at Random): tamamen rassal eksik veri, MAR(Missing at Random): rassal eksik veri ve MNAR(Missing Not at Random): Rassal olmayan eksik veriler şeklinde toplanabilmektedir.

MCAR: veriler eksiklik, tamamen rassal olduğunda eksiklik mekanizması ile ilgili ihmal edilebilen bir durum söz konusu olmaktadır. Bunun nedeni, veri setindeki bir değerin eksik olma nedeninin gözlenen bir değer değerden veya diğer bir değişkene ait değerlerden bağımsız olmasıdır(Allison, 2009). Böyle bir durum var ise eksik verinin MCAR'dır. Anket çalışmalarında soruların eksik olması durumu örnek verilebilir. Bağımsızlık varsayımı çok önemlidir. Buradan hareketle, veri MCAR ise, eksik veriye sahip olmayan değişkenlerin, hedef kitlenin rassal bir örneği olduğunu belirtilir.(Donders ve ark., 2006).

MAR: veri eksikliği, verinin kendi aldığı değerlerden bağımsız iken var olan diğer değişkenlerin kontrol altında olduğu

durumda diğer değişkenlere bağlı olabilme durumunu ifade etmektedir(Schafer,1997). MCAR'a göre nispeten daha zayıf sayılabilecek bir varsayımdır. Soruları okumaktan sıkılan bir anket katılımcısının bazı soruları cevapsız bırakması örnek gösterilebilir.

MNAR: verideki eksik olma durumunun bilinmeyen nedenlerle olduğu, bu sebeple literatürde eksikliğin göz ardı edilemez durumda olduğu eksik veri türüdür. Buradaki durum, verideki eksiklik olasılığı, eksikliğin var olduğu değişkene bağlıdır şeklinde ifade edilebilir(Enders, 2022). Hem MCAR'ın hem de MAR'ın olmadığı durumda MNAR'dan bahsetmek mümkündür(Köse ve Öztemur, 2014). Ankette verilen yanlış bir cevaptan kaynaklı doğrunun çözümlenememesi durumu gösterilebilir.

Eksik veri türleri, bu eksikliğin kapatılmasında önemli bir rol oynamaktadır. Bu türlere uygulanacak yöntem dolayısıyla çıktılar da değişecektir. Bu noktadan bakıldığında, veriyi hazırlama evresindeki ufak hamlelerin bile büyük resimde ne kadar çok noktaya etki edebileceği görülmektedir. Bu sebeple, verinin türünün ya da verilerin arasında bulunan bazı ilişkilerin dikkatle incelenmesi hayati önemdedir(Abidin ve ark., 2018).

2. Makine Öğrenmesi

Değişen ve gelişen teknoloji etkileri elbette bilgi teknolojilerine de yansımıştır. Bu anlamda verinin elde edilmesi, saklanması, taşınması, analize hazır hale getirilip sonucunda bilgiye dönüştürme sürecinde kullanılacak tüm yöntem ve gereçler yeni ihtiyaçlara da cevap verebilir nitelikte olmalıydı. Bu noktada, tanışılan kavramlardan birisi de makine öğrenmesi kavramıdır. Genel olarak, bilgisayarların belirli bir görevi, insan müdahalesinden bağımsız ve doğrudan kodlanmadan yapabilmesine olanak tanıyan bir yapay zekâ dalıdır(Bi ve ark., 2018).Veriler üzerinde istatistiksel ve matematiksel aksiyonları gerçekleştirip veriye ait örüntü ortaya çıkarma ve bu örüntüden hareketle geleceğe yönelik modelleme yapmaya dayanmaktadır(Brynjolfsson ve Mitchell, 2017). Makine öğrenmesini, öğrenme şekillerine göre çeşitli kategorilere ayrılmaktadır.

- Gözetimli öğrenme (Supervised Learning): Bu yöntemde, bilgisayara öğretmek istediğimiz görev ve doğru sonuçları içeren etiketli veriler sunulur. Bu verileri kullanarak, bilgisayar, yeni verileri tahmin etmek için model oluşturur.
- Gözetimsiz öğrenme (Unsupervised Learning): Bu yöntemde, etiketlenmemiş veriler kullanılır ve bilgisayar bu verileri analiz ederek, kalıpları ve yapıları kendisi tespit eder. Bu yöntem daha karmaşık ve öngörülemeyen sonuçlar verebilir.
- Yarı gözetimli öğrenme (Semi-supervised Learning): Bu yöntemde, bir kısmı etiketlenmiş ve bir kısmı etiketlenmemiş veriler kullanılır. Bu yöntem, gözetimli öğrenmenin getirdiği maliyet ve zaman sorunlarına bir çözüm olarak ortaya çıkmıştır.
- Takviyeli öğrenme (Reinforcement Learning): Bu yöntem, bir ajanın belirli bir ortamda öğrenmesine dayanır. Ajan, belirli bir eylemi gerçekleştirerek ortamın mevcut durumunu değiştirir ve bu durumda alacağı ödüle göre hareket eder. Ajanın amacı, ödülü maksimize eden bir strateji geliştirmektir.

Geniş bir çerçeveye sahip olan bu kavram, görüntü işleme, robotik, tıp, dil işleme, finans ve birçok alanda kullanılan makine öğrenmesi aynı zamanda eksik verilerin tamamlanmasında da kullanılmaktadır. Bu anlamda, çalışma kapsamında kullanılan makine öğrenmesi yöntemlerine sonraki bölümlerde yer verilmiştir.

Matched Source

No plagiarism found

Dupli Checker

PLAGIARISM SCAN REPORT

2%
Plagiarised

98%
Unique

Date	2023-05-18
Words	999
Characters	8985

Content Checked For Plagiarism

K-nn, regresyon, sınıflandırma ve eksik veri tamamlama problemlerinin çözümünde kullanılan algoritmalardan biridir. Algoritma, bu çözümlemeleri yapmak için en yakınında bulunan k adet örneğin değerlerini, sınıflarını dikkate alarak operasyonu yapan benzerlik ve mesafe tabanlı bir algoritmadır(Zhang, 2016). Bunun için önce tüm veriler tanımlanır ve bir uzayda gösterilir. Ardından, mesafe olarak en yakın k adet örnek belirlenir. En yakın örnekler belirlenirken kullanılan mesafe ölçümünde Öklid, Manhattan ve Minkowski gibi uzaklık fonksiyonları kullanılmaktadır. Belirlenen örneğin özellikleri tahmin, sınıflandırma, regresyon ve eksik veri tahmini için kullanılır(Mahesh,2020).

Öklid uzaklığı, Pisagor teoreminin 2 boyutlu uzayda uygulanması ile elde edilmektedir. Aşağıda Öklid uzaklığını belirten d(i, j) fonksiyonu ifade edilmiştir.

"d" ("i,j")"=" √(∑_"k=1" ^"p" ▒ ⟦ ⟦"(x" ⟧ _"ik" "-" "x" _"jk" ")" ⟧ ^"2") (1)

Örnekler arası uzaklıkların mutlak değeri şeklinde ifade edilen uzaklık Manhattan uzaklığı olarak tanımlanmaktadır. Aşağıda bu uzaklık fonksiyonu belirtilmiştir.

"d" ("i,j")"=" ∑_"k=1" ^"p" ▒"(" |"x" _"ik" "-" "x" _"jk" | ") i,j=1,2,...,n; k=1,2,...,p" (2)

P-adet değişken için örnekler arası uzaklık hesabında kullanılan bağıntı Minkowski uzaklığıdır ve aşağıdaki gibi tanımlanmaktadır. Aşağıdaki eşitlikte m=1 için Manhattan uzaklığı ve m=2 için Öklid uzaklığı elde edilmektedir.

Random Forest Algoritması(RF)

Makine öğrenmesinde yaygın olarak kullanılan Random Forest algoritması, en basit ifade ile birbirinden bağımsız veya düşük korelasyonlu birçok karar ağacının bir araya getirilmesiyle oluşturulmuş bir karar ormanı topluluğu algoritmasıdır. İlgili karar ağaçları, yine ilgili veri setinden bootstrap yöntemi ile elde edilen örneklerden oluşturulur. Sınıflandırma, regresyon ve eksik veri tamamlama da sık kullanılan algoritmada, her bir karar ağacı tarafından oylanarak ilgili veri kaydının sınıfı belirlenerek operasyon yapılır.

İşleyişinden kısaca bahsetmek gerekirse; RF, öncelikle verileri öğrenmek için birden fazla karar ağacı oluşturur. Bu ağaçlar, verilerin farklı alt kümeleri üzerinde eğitilir. Bu eğitim sırasında, her ağaç rastgele bir özellik alt kümesini kullanarak en iyi karar ağacını oluşturmaya çalışır. Bu nedenle, her ağacın farklı bir bakış açısı ve özellikleri dikkate alması sağlanır. Karar ağaçları eğitildikten sonra, yeni bir veri örneği geldiğinde, her ağaç bu örneğin sınıflandırmasını yapar ve sonuçlar bir oylama yöntemiyle birleştirilir. Bu sayede, her ağacın veri kümesindeki kusurlarının etkisi azaltılır, düşük sapmalar ve yüksek doğruluk oranları sağlar(Tang ve Ishwaran, 2017). Ayrıca, veri kümesindeki gürültüye ve eksik verilere karşı da oldukça dirençlidir.

RF'de m: en iyi bölünmeyi tespit etmek adına düğüm başına kullanılan değişken sayısını, N: geliştirilecek ağaç sayısını göstermek üzere; eğitim veri setinin 2/3 ü ön yükleme örnekleri elde edilir. Eğitim verisinin kalan kısmıyla da hatalar test edilir. Bu 1/3'lük kısma aynı zamanda out of bag verisi de denilmektedir. Her düğümde değişkenler arasından m değişkenleri rassal olarak seçilerek değişkenler arasından en iyi dal tespit edilir. Sınıflandırma uygulamalarında, her yaprak sadece bir sınıfa giren sınıf elemanlarını içerecek şekilde oluşturulur. Regresyon uygulamalarında ise ilgili yaprakta çok az sayıda birim kalana kadar ilgili ağaçlar bölünmeye devam etmektedir.

Amelia

Amelia, eksik veri tamamlamada kullanım alanına sahip istatistiksel bir modelleme yapısıdır. Bu yapı, eksik verileri

tamamlarken otomatik olarak çoklu tamamlama yöntemleri kullanmaktadır. Eksik olan verilerin doğal dağılımının analizi ile verilerin eksik olma olasılığını ortaya çıkarmaktadır. Böylece ilgili bilgilerle istatistiksel olarak uygun yöntem ile eksik veriler tamamlanmaktadır. Amelia algoritması, çok boyutlu veri yapılarında kullanılabilen farklı tamamlama yöntemlerini birleştiren bir yapıya sahiptir. İyi sonuçlar üretmesinin yanında araştırmacının, istatistiksel modelleme konusunda uzman olmasını gerektirmediği için ekstra bir kullanım sıklığına da sahiptir.

Amelia algoritması, bootstrap ve beklenti maksimizasyonu(EM) algoritmalarını kullanarak çoklu tamamlama yapmaktadır. Algoritma, eksik olan verilerin tamamlanması için birden fazla tamamlama modeli uygular ve her bir modelde bootstrap ve EM algoritmalarını kullanarak tamamlanan eksik verilerin ortalamasına alarak elde edilen sonuçları birleştirmektedir(Honaker ve ark., 2011).

Bootstrap kısaca, ilgili verilerin örneklem çekme yoluyla yeniden örnekleme yaparak örneklem dağılımı hakkında çıkarımların yapılmasını imkân veren bir yöntemdir(Doğan,2017). Amelia algoritması da, farklı örneklemlerden elde edilen birden çok eksik veri setlerini oluşturabilmek için kullanmaktadır. EM algoritması ise kısaca, iteratif olarak eksik veri yerine tamamlanacak değerlerin güncellenerek en yüksek olabilirlik tahminlerini bulmak adına kullanılmaktadır(Doğru ve ark.,2016). Böylece Amelia, EM yoluyla her bir eksik veri seti için eksik verileri tamamlar ve tamamlanan değerleri kullanarak veri setinin tamamında eksik değerlerin ortalamasını almaktadır. Amelia algoritması, bu şekilde bootstrap ve EM algoritmalarının kombinasyonları ile iyi sonuçlar sunan ve güvenilir eksik veri tamamlama sonuçları sunmaktadır.

Stokastik Regresyon(SR)

Stokastik regresyon, istatistiksel bir regresyon modeli türüdür. Bu model, özellikle zaman serisi verileri gibi sürekli bir değişkenin geçmiş değerleri ile gelecekteki değerleri arasındaki ilişkiyi analiz etmek için kullanılır. Birçok değişkenden etkilenen bir sürekli değişkenin tahmininde kullanılan bir modeldir. Bu modelde, sürekli değişkenin tüm değişkenlere bağlı olduğu kabul edilir ve bu bağımlılık stokastik bir şekilde modellenir. Stokastik regresyon modelleri, genellikle Gauss markov varsayımlarının geçerli olduğu ve gürültülü verilerin olduğu durumlarda kullanılır. Bu modelde, sürekli değişkenin gelecekteki değerleri, rastgele bir hata terimi ve bir dizi bağımsız değişkenin doğrusal kombinasyonu ile açıklanır. Bu modelin en önemli özelliklerinden biri, hata teriminin normal dağılıma sahip olduğu ve tüm hata terimlerinin birbirinden bağımsız olduğu varsayımıdır. Bu varsayımlar altında, stokastik regresyon modeli, maksimum olabilirlik yöntemi ile tahmin edilir ve gelecekteki değerler için tahmin aralıkları belirlenebilir.

SR, eksik veri tamamlamada tercih edilen yöntemlerden biridir. SR ile eksik veri tamamlama, lineer regresyon doğrusundaki değerlerin eksik veri yerine kullanılmasından kaynaklı veri dağılımının etkilenmesi ve kovaryans değerlerinin daha düşük kestirilmesi benzeri problemleri çözümlemek adına geliştirilmiştir. Bu şekilde elde edilmiş değerlerle kurulmuş olan regresyona sıfır ortalamalı normal dağılama sahip hatalar eklenmektedir. Enders(2010), bu yöntemin, rassal ve tamamen rassal şekilde eksiklik taşıyan veri düzenlerinde yansız parametre tahminleri ve lineer regresyon ile eksik veri tamamlama yöntemine göre çok daha iyi sonuçlar ürettiğini ifade etmiştir(Baraldi ve Enders,2010).

Ortalama Atama

Ortalama atama yöntemi, eksik verileri, veri setindeki ilgili değişkenin var olan değerlerine ait ortalamayı eksik olan değer yerine koyarak tamamlamaktadır. Literatürde, eksik veri tamamlama uygulamalarında oldukça sık bir kullanıma sahiptir.

Naive Bayes Algoritması

İstatistiksel sınıflandırma araçlarından biri olan ve veri madenciliği uygulamalarında yaygın bir şekilde kullanılan naive bayes sınıflandırma algoritması, sınıf nitelik değeri araştırılan bir veri kümesinin, her bir sınıf niteliğine sahip olma olasılıklarının hesaplanıp, hesaplanan olasılıklardan en büyük değere sahip olan nitelik değerini, sınıf niteliği değeri olarak atayan sınıflandırma aracıdır. Naive Bayes, önceden sınıflanmış vaziyette olan verileri kullanarak, yeni veri kayıtlarının, tanımlanmış her sınıf değerini alma olasılıklarını hesaplamaktadır(Dondurmacı,2011). Veri kümelerinin içerdiği nitelik değerlerinin koşullu olarak birbirinden bağımsız olduğu varsayımı, yapılacak sınıflandırma tahminlerinin analiz sürecini kolaylaştırmaktadır(Vembandasamy ve ark., 2015).

Matched Source

Similarity 10%

Title:Issue Full File - Food and Health - ScientificWebJournals

1% and 5% citrus fiber was not significant (p>0.05) in com- parison to change in the pH of the ... ODS-2 column (5 µm, 125 x 4 mm; Waters, Milford, MA,.

http://jfhs.scientificwebjournals.com/tr/download/issue-full-file/46056

Dupli Checker

PLAGIARISM SCAN REPORT

2%
Plagiarised

98%
Unique

Date	2023-05-18
Words	816
Characters	7788

Content Checked For Plagiarism

Kullanılan algoritmaların performanslarını değerlendirmek kıyaslamak amacıyla bazı kriterler kullanılmaktadır. Bu kriterler, sınıflandırma süresi, doğruluk, kesinlik, duyarlılık, F-ölçütü ve ROC(Receiver Operating Characteristic) alanı değerleridir. Sınıflamada kullanılan algoritmaların başarısı, doğru ve yanlış sınıf değeri atanmış örneklerle bağlantılıdır. "Confusion Matrix" adı verilen matris yapısı bu değerlere ait çıktıyı sunmaktadır. Tablo-1'de verilen bu matris değerleri ile performans kriterleri hesaplanmaktadır. Bu matriste, x; TP= True Pozitif, y; FN= False negatif, z; FP= False Pozitif ve k; TN=True Negatif değerlerini ifade etmektedir(Oprea ve Ti,2014).

Doğruluk oranı; doğru sınıfa atanmış örnek adedinin toplam örnek adedine bölümü ile bulunur. Bu değerden, sınıflandırma sonucunun gerçeğe ne kadar yakın olduğunu göstermektedir. Bu ölçüt, sınıflandırma sonucunun gerçek değerlere ne kadar yakın olduğunu gösteren bir orandır.

Kesinlik; sınıf 1 olarak sınıflandırılmış TP örnek adedinin, sınıf 1 olarak sınıflandırılmış tüm TP ve FP örnek sayısına bölünmesiyle bulunur. Kesinlik oranının araştırılmasının sebebi, aynı yöntemle elde edilmiş analiz sonuçlarının birbirine yakınlığının tespit edilmeye çalışılmasıdır.

Duyarlılık; TP olarak sınıflandırılmış örnek sayısının toplam pozitif TP ve FN toplamlarına oranıdır. Sınıflandırma işlemi sonuçlarının birbirine olan yakınlığının bir göstergesidir.

Duyarlılık ve kesinlik değerlerinin harmonik ortalaması F-ölçütü kriterini tanımlanmaktadır. Duyarlılık ve kesinlik geçerli ve değerlendirme anlamında önemli kriterler olsa da tek başlarına yeterli
kıyaslama kalitesinde olmayabilirler. F-ölçütü, dengeli bir duyarlılık ve kesinlik durumunda yüksek değerler almaktadır.

ROC alanı; TP oranının Y ekseninde ve FP oranının X ekseninde karşılaştırıldığı 2-boyutlu bir grafiğin oluşturduğu alandır.

Daha büyük alan değeri daha iyi performans anlamına gelmektedir. Bu alan, veri madenciliği algoritmalarının karşılaştırmasında genel bir kullanıma sahip bir kriterdir

Çalışma kapsamında makine öğrenmesi ve veri bilim çalışmalarında sık kullanıma sahip olan "Hitters" veri seti kullanılmıştır. Bu veri seti, 1986 ve 1987 yıllarında Major Ligi'ndeki beyzbol sporcularına ait oyun istatistikleri ve yıllık aldıkları ücret(bin dolar) verilerinden oluşmaktadır. Ayrıca veri setinde oyuncuların aldıkları yıllık ücretler dikkate alınarak en çok olan ücretten en düşük ücrete doğru A, B, C, D ve E şeklinde kodlanıp kategorize edilerek sınıflandırılmış veri haline getirilmiş ve "Player" değişkeni düzenlenmiştir. Böylece, ilgili veri seti, 18 adet değişken ve değişkenlere ait 322 adet kayıttan oluşmaktadır.

Veri setindeki veriler manipüle edilerek verilerin %5'i eksiltilmiştir. Manipülasyon işlemi R programlama dili kullanarak yapılmıştır. Eksiltilmiş olan veriler, ilgili değişken değerlerinin ortalamasıyla, stokastik regresyon yöntemiyle yanı sıra daha önce bahsedilmiş olan K-nn, Random Forest ve Amelia makine öğrenmesi algoritmalarıyla tamamlanmıştır. Ardından, bu algoritmalarla tamamlanarak elde edilen her bir veri setine ve veri setinin manipüle edilmemiş orijinal haline Naive Bayes algoritmasıyla "Player" değişkeni sınıf nitelik olmak üzere WEKA programı kullanılarak sınıflandırma operasyonu yapılmıştır. Böylece, temel yöntemlerin en başında gelen ortalama atama ile makine öğrenmesi algoritmalarının, eksik veri tamamlama performansları karşılaştırılmış aynı zamanda veri madenciliği ve makine öğrenmesinde sık başvurulan sınıflandırma analizine olan etkileri ortaya koyulmuştur.

İlk olarak veri setin manipüle edilmeden önce orijinal hali Naive Bayes algoritmasıyla "Player" değişkeni sınıf niteliği olmak üzere sınıflandırılmıştır. Sonuçlar Şekil-2'de gösterilmiştir.

Ardından manipüle edilen veri seti, daha önce bahsedilen yöntemlerle tamamlanmış ve yine Naive Bayes algoritmasıyla

sınıflandırılmıştır. Tüm sınıflandırma sonuçları özet olarak Tablo-3'te verilmiştir. Sınıflandırma sonuçlarına ait çıktılar Ekler bölümünde sunulmuştur.

Bu çalışmada, tüm sınıflandırma operasyonları Naive Bayes algoritması ile yapıldığı için sınıflandırma sonuçları üzerinde farkı oluşturan unsur sadece eksik verileri tamamlama algoritmalarıdır. Yapılan sınıflandırma sonuçlarına bakıldığında, makine öğrenmesi algoritmalarına ait performans değerlerinin, veri setinin manipüle edilmemiş haline ait sınıflandırma performans değerlerine anlamlı derecede yakın olduğu görülmektedir. Bu yakınlığın, tüm performans değerlendirme kriterleri için geçerli olduğu ortadadır.

Ortalama atama yöntemine ait performans değerleri incelendiğinde, veri setinin manipüle edilmemiş haline uygulanmış sınıflandırma işlemine ait performans değerlerinin gerisinde kaldığı gözlenmektedir. Bunun aksine makine öğrenmesi algoritmalarına ilişkin performans sonuçları, orijinal veri setinin performans değerlerine oldukça yakın olmasıyla birlikte ortalama atama yöntemine ait performans değerlerinden anlamlı derecede daha üstündür. Sınıflandırma süresi kriterine göre makine öğrenmesi yöntemlerinin, veri setinin manipüle edilmemiş haline dayalı durumdan genel olarak daha kısa sürede sınıflandırma operasyonu yaptığı gözlenmiştir. Çok büyük hacimli veri setleri ile çalışıldığında kısa sürede aynı operasyonu yapabilmek oldukça öne çıkan ve aranan bir özelliktir. Böylece, kısa sürede sınıflandırma yapabilme özelliği bakımından makine öğrenmesi yöntemlerinin ön planda yer aldığı görülmüştür. Doğruluk, kesinlik, duyarlılık, F-ölçütü ve ROC alanı kriterlerinde de, makine öğrenmesi algoritmalarının performanslarına ait değerlerin, orijinal durumunun performans değerlerine oldukça yakın seviyede seyrettiği yanı sıra özellikle K-nn ve Random Forest algoritmalarının genel olarak çok üstün bir başarı gösterdiği tüm kriterlerce gözlenebilmektedir. Hatta K-nn algoritmasının, veri setinin manipüle edilmiş haline ait performans değerlerimden daha iyi performans değerlerine ulaştığı ortadadır. Bu durum, K-nn algoritmasının tamamladığı eksik değerlerin, veri setinin yapısına ve karakteristiğine çok iyi uyum sağlayan değerler olduğu böylece, sınıflandırma operasyonunda da başarıyı yükselten bir etki yaptığı görülmektedir.

Çıktılara bakıldığında, makine öğrenmesi ile yapılan eksik veri tamamlama yöntemlerine ait performans değerleri birbirlerine yakın olsalar da, kendi aralarında ufak farklılıklar içermektedir. Sonuçlara göre, makine öğrenmesi algoritmaları arasında en iyi performansı gösteren algoritmanın K-nn olduğu ve Amelia algoritmasının, diğer makine öğrenmesi algoritmalarına nispeten daha düşük performans gösterdiği tüm kriterlerce gözlenmiştir. Böylece, makine öğrenmesi algoritmalarının eksik veri tamamlamada oldukça başarılı olduğu görülmüştür. Ayrıca, K-nn algoritmasının çalışma kapsamında elde ettiği sonuçlar gibi makine öğrenmesi algoritmalarının, sınıflandırma performanslarına katkı yaparak olumlu bir etki yapabileceği de gözlenmiştir.