Journal of the Faculty of Engineering and Architecture of Gazi University 33:1 (2018) 323-336

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University

Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

# Sınıflandırma için diferansiyel mahremiyete dayalı öznitelik seçimi

Esra Var[1], Ali İnan[2]*

[1]Işık Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul, 34980, Türkiye
[2]Adana Bilim ve Teknoloji Üniversitesi, Mühendislik Fakültesi, Bilgisayar, Mühendisliği Bölümü, Adana, 01100, Türkiye

## Ö N E Ç I K A N L A R

- Diferansiyel mahremiyet yöntemi ile gizlilik koruması sağlanması
- Bilinen ilk diferansiyel mahremiyet ile öznitelik seçimi yöntemi
- WEKA veri madenciliği kütüphanesi ile entegre edilmiş, açık kaynak kodu

Makale Bilgileri
Geliş: 30.03.2017
Kabul: 13.08.2017

DOI:
10.17341/gazimmfd.406804

Anahtar Kelimeler:
Diferansiyel mahremiyet, sınıflandırma, öznitelik seçimi

ÖZET

Veri madenciliği ve makine öğrenmesi çözümlerinin en önemli ön aşamalarından biri yapılacak analizde kullanılacak verinin özniteliklerinin uygun bir alt kümesini belirlemektir. Sınıflandırma yöntemleri için bu işlem, bir özniteliğin sınıf niteliği ile ne oranda ilişkili olduğuna bakılarak yapılır. Kişisel gizliliği koruyan pek çok sınıflandırma çözümü bulunmaktadır. Ancak bu yöntemler için öznitelik seçimi yapan çözümler geliştirilmemiştir. Bu çalışmada, istatistiksel veritabanı güvenliğinde bilinen en kapsamlı ve güvenli çözüm olan diferansiyel mahremiyete dayalı özgün öznitelik seçimi yöntemleri sunulmaktadır. Önerilen bu yöntemler, yaygın olarak kullanılan bir veri madenciliği kütüphanesi olan WEKA ile entegre edilmiş ve deney sonuçları ile önerilen çözümlerin sınıflandırma başarımına olumlu etkileri gösterilmiştir.

# Differentially private attribute selection for classification

## H I G H L I G H T S

- Privacy protection through differential privacy
- First known method to propose differentially private attribute selection
- Public source code integrated with the WEKA data mining library

Article Info
Received: 30.03.2017
Accepted: 13.08.2017

DOI:
10.17341/gazimmfd.406804

Keywords:
Differential privacy, classification, attribute selection

ABSTRACT

Selecting a relevant subset of attributes is one of the most important data preprocessing steps of data mining and machine learning solutions. For the classification task, selection is based on the correlation between an attribute and the class attribute. There are various studies on privacy preserving classification. However, there is no attribute selection solution for such work in the literature. In this study, novel attribute selection methods based on the state of the art solution in statistical database security, known as differential privacy, are proposed. The proposed solutions are implemented with the popular data mining library WEKA and experimental results confirm the positive effects of the proposed solutions on classification accuracy.

*Sorumlu Yazar/Corresponding Author: ainan@adanabtu.edu.tr / Tel: +90 322 455 0000 / 2144

## 1. GİRİŞ (INTRODUCTION)

Bilgisayar kullanımının giderek yaygınlaşması, pek çok sosyal ve ticari faaliyetin elektronik ortamda gerçekleşmeye başlamasının doğal bir sonucu olarak, gerek özel teşebbüsler, gerekse kamu kurum ve kuruluşları bireylere ait pek çok veriyi depolar hale gelmiştir. Bireylerin kişisel gizliliklerinin ihlal edilebileceğine yönelik kaygılarının bir sonucu olarak hem toplanan verinin işlenmesini sağlayacak, hem de kişisel gizliliğin ihlal edilmesine engel olacak yöntemler geliştirilmeye başlanmıştır.

Kişisel gizliliği korumayı hedefleyen bu çalışmaları iki kategoride incelemek mümkündür:

- Yatayda (kayıtlar üzerine) [1] veya dikeyde (öznitelikler üzerine) [2] dağıtık olan veri kümelerinin üzerindeki hesaplamaları şifrelemeye dayalı yöntemler ile gerçekleştiren çalışmalar
- Dağıtık olmayan veri kümelerinde yer alan kayıtların kişilerle bağlantılarını kopartmaya çalışan ve bu "anonimleştirme" işleminin ardından veri kümesinin açık olarak paylaşılabileceğini varsayan k-anonimleştirme [3], l-çeşitlilik [4], t-yakınlık [5] ve benzeri yaklaşımlar.

Her iki kategorinin de kendine özgü zayıf yanları bulunmaktadır. Şifrelemeye dayalı çözümler sadece dağıtık veri senaryolarına uyarlanabilir ve içerdiği yoğun hesaplamalı işlemlerden dolayı ağır çalışır. Buna karşılık anonimleştirme yöntemleri kişisel gizliliği korurken sadece belirli bir anonimlik tanımının ifade ettiği saldırgan modellerine karşı güvence sağlayabilir. "Diferansiyel mahremiyet" [6] adlı yeni bir çözüm, veri kümesine erişimi istatistiksel bir sorgulama arayüzüne indirger. İstatistiksel olmayan, doğrudan kayıt temin etmeyi amaçlayan tüm sorgular engellenir. İstatistiksel sorguların cevaplarına ise sorguların "hassasiyetine" uygun olarak gürültü eklenir. Bu yöntem, erişim hakkını kullanarak veri kümesinde kaydı bulunan şahısların kişisel gizliliğini ihlal etmeye çalışan olası tüm saldırganları göz önüne almaktadır. Saldırganı herhangi bir arka plan bilgisine kısıtlı görmediği için çok güvenli bir istatistiksel veritabanı koruma mekanizması olduğu değerlendirilmektedir. Bu makalede, diferansiyel mahremiyet kullanılarak öznitelik seçiminin nasıl yapılabileceği tartışılmakta ve ortaya konulan çözüm yöntemlerinin sınıflandırma işlemindeki başarımı yapılan detaylı deneyler ile ortaya konulmaktadır. Öznitelik seçimi, veri madenciliği ve makine öğrenmesi çözümlerinin bir ön aşamasıdır. Kullanılacak veri kümesi içinde yer alan özniteliklerin, yapılacak işlem için en uygun olan alt kümesinin seçilmesine "öznitelik seçimi" denir. Sınıflandırma işlemi için bu seçim, öğrenme verisinin incelenmesi yoluyla yapılır ve her özniteliğe sınıf özniteliği ile ne oranda ilişkili olduğunu gösteren bir değer atanmasına dayanır. Filtrelemeye dayalı öznitelik seçimi yöntemleri için bilgi kazanımı ve ki-kare en etkili değerleme ölçütleri arasındadır [7]. Bu sebeple bu çalışmada, bir özniteliğin bilgi kazanımı ve ki-kare değerlemesinin yapılmasını sağlayacak veri erişimi SQL sorguları cinsinden ifade edilmiştir. Bu sorgu kümesinin diferansiyel mahremiyete göre tanımlanmış hassasiyet değeri hesaplanmış ve öğrenme verisinden elde edilen doğru cevaplara uygun miktarda gürültü eklenerek öznitelik seçimi yöntemleri uygulanmıştır. Çalışmanın katkıları kısaca şu şekilde özetlenebilir:

- Diferansiyel mahremiyete dayalı gizlilik korumasının nasıl çalıştığının ilk kez teorik olarak ve pratik örneklerle Türkçe anlatımı,
- Öznitelik seçimi problemi için bilinen en popüler yöntemler olan ki-kare ve bilgi kazanımının diferansiyel mahremiyette kullanılan hassasiyet değerlerinin hesaplanması,
- Diferansiyel mahremiyet ile korunmuş veri kümesi üzerinde bilinen ilk öznitelik seçimi yöntemi olması,
- Detaylı deney senaryoları ile pek çok farklı (az sayıda öznitelik içeren, az sayıda kayıt içeren, vs.) veri kümesi üzerinde tüm olası parametrelerin etkilerinin incelenmesi ve tartışılması,
- Çözümlerin WEKA [8] adlı popüler veri madenciliği kütüphanesi ile entegre edilmesi ve ilgili kodlara açık erişim sağlanması.

Makalenin devamı şu şekildedir. Bölüm 2'de bu çalışma ile ilgili var olan çalışmalar tartışılmaktadır. Bölüm 3'te diferansiyel mahremiyet ve Bölüm 4'te öznitelik seçimi konuları ana hatları ile anlatılmaktadır. Bölüm 5'te önerilen çözüme ve Bölüm 6'da deney sonuçlarına yer verilmiştir. Bölüm 7'de ise sonuçlar özetlenmektedir.

## 2. İLGİLİ ÇALIŞMALAR (RELATED WORK)

İstatistiksel veritabanı güvenliği konusunda pek çok çalışma yapılmıştır. Veri temizleme çözümleri bunlardan bir tanesidir. Sweeney [3] bir veri setinde yer alan kayıtların genelleştirme ve bastırma araçları ile k-anonimleştirilmesini çalışmıştır. Aggarwal veri setinin çok fazla öznitelik içermesi halinde k-anonimleştirmenin etkin olarak çalışmadığını göstermiştir [9]. Machanavajjhale vd. [4] k-anonimlik tanımı üzerine olası saldırıları aktarır ve hassas niteliklerde birden çok değerin "yeterli temsil"ine dayalı l-çeşitlilik tanımını önerir. L-çeşitlilik tanımının da belirli saldırganları yeterince modellemediğini gösteren Li vd. [5] t-yakınlık adlı anonimlik tanımını yapar. Dwork vd. bütün bu tanımların teorik olarak yetersiz olduğunu ispat eder [6]. Buna göre, olası her anonimlik tanımı için başarılı olacak bir saldırgan mevcuttur. Güvenlik mekanizması, kişilerin veritabanına katılma kararından en az seviyede etkilenmesini temin etmelidir. Bir veritabanı dışarı erişime açıldığı sürece, belirli gizli bilgilerin ifşa olması kaçınılmazdır. Diferansiyel mahremiyet tanımı oldukça güçlü bir koruma sağlamaktadır ve hızlı bir şekilde popülerlik kazanır. Zou vd. [10] büyük veri kümelerinde veri değiştirme işleminin çalışma süresini kısaltan bir yöntem sunar. Evans vd.'nin çalışması [11] mikro veriye gürültü eklenmesi ile ilgilidir. Okkalıoğlu vd. [12] ortak filtreleme yöntemlerinin dağıtık veritabanlarına uygulanmasına yönelik saldırıların gizlilik tehditlerini ortaya

koyar. Shlomo vd. [13] diferansiyel mahremiyet kullanarak alınacak bir örnekleme gürültü eklenmesi problemini inceler. Kadampur vd. [14] karar ağacı analizinin kişisel gizliliğe zarar vermeyecek şekilde gürültü eklenerek gerçekleştirilmesini sağlayacak yöntemler sunar.

Soria-Comas vd. [15] diferansiyel mahremiyeti farklı bir açıdan ele almaktadır. Çalışmalarında çok değişkenli sorgulamalar için daha iyi sonuç verecek ve diferansiyel mahremiyet şartlarını sağlayan, veriden bağımsız, optimal bir gürültü dağılımı belirlemektedirler. Diferansiyel mahremiyet çözümü pek çok farklı alanda uygulanmaktadır. Guang vd. [16] PINQ altyapısını kullanarak, hasta veri kümelerinde daha kolay ve hasta gizliliğini ihlal etmeyen analizler yapılabileceğini gösterir. Lee vd. [17] mobil sağlık hizmetleri için gizliliği koruyan ve öznitelik seçimi içeren bir çözüm sunmaktadır. Bu çözüm, bu çalışmadan farklı olarak bilgi paylaşma paradoksuna dayalı olarak hangi özniteliklerin paylaşılacağına dikkatle karar verilmesi gerektiğini belirtmektedir.

Divanis vd. [18] elektronik sağlık verisinin paylaşımı için kullanılabilecek gizliliği koruyan algoritmaların bir taramasını vermektedir. Öznitelik seçimi yöntemleri filtreleme çözümleri, kapsama çözümleri ve gömülü çözümler olarak kategorilere ayırılabilir. Bunlar arasında en yaygın kullanılan filtreleme çözümleridir. Bu çalışmada filtreleme çözümleri ile öznitelik seçimi araştırılmaktadır. Filtreleme çözümleri her öznitelik için değerleme yapar ve değer sırasına göre öznitelik seçer. Yang vd. [7] filtreleme çözümleri için bilgi kazanımı ve ki-karenin en etkili yöntemler olduğunu göstermiştir. Öznitelik seçiminde karşılıklı bilgiye dayalı yöntemler Çelik vd. [19] tarafından incelenmiş ve yeni bir sezgisel fonksiyon önerilmiştir. Akben vd. [20] Parzen pencereleme kullanarak öznitelikler değerlerinin yoğunluk katsayısına dönüştürülmesinin sınıflandırma başarımını %18'e kadar artırdığını göstermiştir.

Ancak bu çalışmaların hiçbiri öznitelik seçimi esnasında mahremiyet korumasını göz önünde bulundurmaz. Diferansiyel mahremiyet kullanan pek çok veri madenciliği yöntemi bulunmaktadır. Öznitelik seçimine yönelik pek çok yöntem önerilmiştir. Ancak diferansiyel mahremiyete dayalı öznitelik seçimi çözümü bilinmemektedir. Bu sebeple, önerilen çözümler özgün çözümlerdir.

## 3. DİFERANSİYEL MAHREMİYET (DIFFERENTIAL PRIVACY)

Diferansiyel mahremiyet yöntemi, korunacak veritabanına doğrudan erişimi engeller. Bunun yerine, veritabanı üzerinde istatistiksel bir sorgulama arayüzü sağlar. Temel mekanizma, kullanıcının veritabanına bir sorgulama betiği göndermesi, diferansiyel mahremiyet ile korunan veritabanının bu sorguların sonuçlarına gürültü eklemesidir. İstatistiksel olmayan, cevabı belirli bir kayıt kümesi olan sorgulara cevap verilmez. Doğrudan kayıt erişimi bu şekilde engellenir. Şekil 1, bu çalışma mekanizmasını görsel olarak özetlemektedir. Sorgu cevaplarına eklenecek gürültü miktarı, sorgu betiğinin "hassasiyet" değeri ile ilgilidir. Bir betiğin hassasiyeti, aralarında tek bir kayıtlık fark olan herhangi iki "kardeş" veritabanı üzerinden hesaplanır. Hassasiyet tanımı, tüm kardeşler üzerinde alınabilecek sorgu cevapları arasındaki en yüksek L1 mesafe şeklindedir. Hassasiyetin hesaplanması NP-zor bir işlemdir [21].

Tanım 3.1. (Hassasiyet) $Q$ ile gösterilen bir sorgu betiğinin hassasiyeti $\Delta$ şu şekilde hesaplanır:

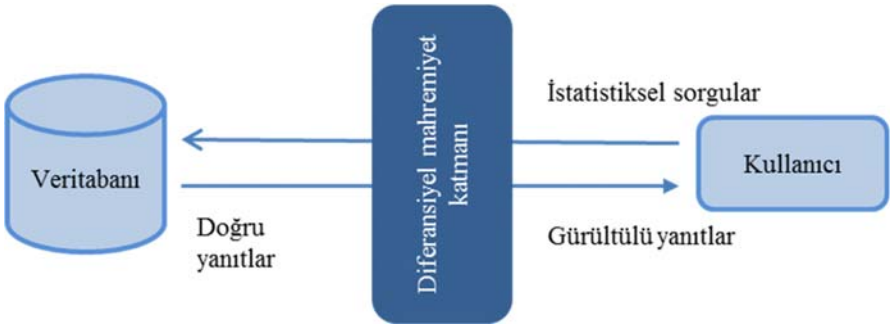
$$\Delta = \underset{\substack{kardeş\ veri \\ tabanları\ D,D'}}{\text{argmax}} \; || S^D(\text{Q}) - S^{D'}(\text{Q}) ||_1 \qquad (1)$$

Eş. 1'de $S^D(Q)$ (veya $S^{D'}(Q)$) ile ifade edilen, Q sorgu kümesinin $D$ (veya $D'$) üzerindeki cevap vektörüdür. Tüm olası kardeş veritabanları $D$ ve $D'$ için L1 normdaki vektör uzaklığının en büyük değeri, $Q$'nun hassasiyet değeridir. Hassasiyet, tanımı gereği $Q$'nun uygulanacağı veritabanından bağımsızdır [22]. Hassasiyet, ilgili kayıtlar üzerinde ne kadar detaylı bilgi sorgulandığını ölçer. Eğer, kardeş veritabanı tanımına göre, bir kayıtlık bir değişiklik $Q$'nun cevabını çok fazla etkiliyorsa, hassasiyet hesabı yüksek çıkacaktır.

*Örnek:* Sorgu kümesi $Q$ içinde aşağıdaki sorgular bulunsun:

- *Q1: SELECT COUNT(*) FROM T WHERE CİNSİYET = 'M'*
- *Q2: SELECT COUNT(*) FROM T WHERE YAŞ > 40*

$D$ ve $D'$ kardeş veritabanları arasında tek kayıtlık fark olacaktır. Bunların $D = \{r_1, r_2, ..., r_k\}$ ve $D' = \{r_1, r_1, ..., r'_k\}$ oldukları varsayılsın. Şekil 2'de gösterildiği üzere, $r_k$ ve $r'_k$

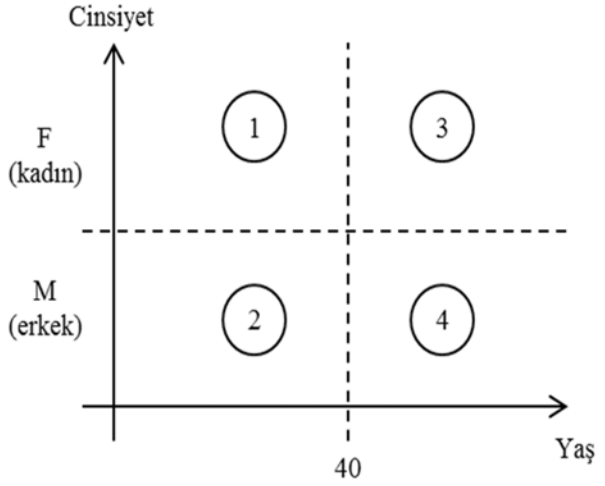



**Şekil 1.** Diferansiyel mahremiyetin kullanımı (Usage of differential privacy)

kayıtlarının yer aldığı olası bölgelere göre 16 farklı durum oluşması mümkündür. Bazı durumlar incelenirse:

- Hem $r_k$, hem $r'_k$ aynı bölgedeyse: cevaplar aynı olur.
- $r_k$ bölge 1, $r'_k$ bölge 2'de ise: Q1 cevapları 1 farklı olur, Q2 cevapları aynı olur. Toplam fark 1.
- $r_k$ bölge 1, $r'_k$ bölge 3'de ise: Q1 cevapları aynı olur, Q2 cevapları 1 farklı olur. Toplam fark 1.
- $r_k$ bölge 1, $r'_k$ bölge 4'de ise: Q1 ve Q2 cevapları birer farklı olur. Toplam fark 2.

Tüm durumlar incelenirse görülecektir ki, örnek sorgu betiği $Q$ için olası en yüksek fark 2'dir.




**Şekil 2.** Örnek sorgu kümesi için kayıt uzayının bölgeleri
(Regions of the record space for the sample query set)

Tanım 3.2. (Diferansiyel mahremiyet) $D$ ve $D'$ kardeş veri kümeleri, $R$ olası cevap uzayı olmak üzere, rastsallaştırılmış bir algoritma $K$ ε-diferansiyel mahremiyeti eş. 2 ile sağlar.

$$Pr[K(D) \in R] \leq e^{\varepsilon} \times Pr[K(D') \in R] \quad (2)$$

Bu tanımı sağlayan pek çok mekanizma geliştirilmiştir. Bu çalışmada Dwork vd.'nin [6] önerdiği ilk ve popüler mekanizma olan Laplace mekanizması kullanılmıştır. Buna göre, sorgu kümesi hassasiyeti Δ belirlendikten sonra, her sorgu cevabına bağımsız olarak Laplace dağılımından seçilen ortalaması 0, şiddeti ise en az Δ/ε olan gürültü eklenir. Hassasiyet yüksek olursa veya ε düşük olursa gürültü miktarı artmaktadır.

## 4. ÖZNİTELİK SEÇİMİ (ATTRIBUTE SELECTION)

Öznitelik seçimi belirli bir öznitelik uzayı içerisinden yapılacak analiz ile ilgili en yüksek faydayı sağlayacak öznitelik alt kümesinin belirlenmesini sağlayan yöntemlerin genel adıdır. $n$ öznitelikli bir veri seti için bu uzayın büyüklüğü $2^n$ olacaktır. Veri seti içerisinde yer alan bütün özniteliklerin kullanılması gereksiz hatta zararlı olabilir. Sınıflandırma işlemi için gereksiz öznitelik kavramı, sınıf değeri ile hiçbir bağıntısı bulunmayan öznitelikleri ifade eder. Örneğin, bir arabanın yakıt tüketimi ile rengi arasında hiçbir ilişki bulunmaması beklenen bir durumdur. Bu özniteliklerin veri setinde tutulması, geliştirilecek modelin karmaşıklaşmasına sebep olabilir. Aralarındaki korelasyon yüksek olan özniteliklerin (ör. doğum tarihi ve yaş) bir arada kullanılması ise ilgili öznitelik çiftine gereğinden fazla ağırlık verilmesi anlamına gelir ve sınıflandırma yöntemini yanlış yönlendirir. Bu sebeplerle öznitelik seçimi pek çok makine öğrenmesi ve veri madenciliği yöntemi için hayati öneme sahiptir. Aşağıda öznitelik seçimi için yaygın olarak kullanılan yöntemler özetle aktarılmaktadır.

### *4.1. Filtreleme Çözümleri (Filtering Solutions)*

Bu çözümlerde her özniteliğe bir değer atanır ve öznitelikler atanan değerlere göre sıralanır. Bu aşamada hangi özniteliklerin seçileceği (dolayısıyla hangilerinin çıkartılacağı) bir filtre yardımıyla belirlenir. Yaygın olarak kullanılan filtreler, listenin ilk $n$ elemanını seçme veya atanan değeri $t$ eşiğinin üzerinde olan öznitelikleri seçme şeklindedir. Eşik belirleme işlemi veri setine bağlı olacağı için, öznitelik seçimi çalışmaları daha ziyade seçilecek öznitelik sayısı üzerine deneyler sunmaktadır. Özniteliklere değer atamak amacıyla yaygın olarak kullanılan ölçütler arasında ki-kare, bilgi kazanımı ve korelasyon katsayısı bulunmaktadır. Her ne kadar $n$ tane öznitelik için çözüm uzayı $2^n$ boyutlu olsa da, filtreleme çözümleri yaygın olarak açgözlü algoritmalar olarak kullanılır ve (her nitelik alt kümesi yerine) her niteliğe tek bir değer atanarak uzay $n$ boyuta indirgenir.

### *4.2. Kapsama Çözümleri (Wrapper Solutions)*

Kapsama çözümleri $2^n$ elemanlı öznitelik alt-kümeleri uzayının tamamını üretir ve değerlendirir. Seçenekler arasındaki tercih ise öznitelik seçimi sonrasında kullanılacak olan öğrenme yönteminin kalite ölçütüdür. Öğrenme yöntemi "kara kutu" şeklinde kullanılır [23]. Örneğin, sınıflandırma için doğruluk gibi bir ölçütten faydalanılmaktadır. Özyineli öznitelik eleme algoritması kapsama çözümlerinde yaygın olarak kullanılan bir çözümdür. Kapsama çözümlerine dayalı öznitelik seçiminin sınıflandırma yönteminden bağımsız olması mümkün değildir. Bu durum ise, diferansiyel mahremiyet kullanılmasını zorlaştırmaktadır. Sınıflandırma yöntemini öznitelik seçimi yöntemi ile birleştirmek, birden çok model üretmekle denktir ve gürültü miktarını çok fazla arttırarak anlamlı bir sınıflandırma modeli oluşturulmasına engel olacaktır.

Kapsama çözümlerinin olası diğer dezavantajları ise şu şekildedir:

- Kayıt sayısının az olması durumda modelin öğrenme veri kümesine aşırı uyum sağlaması,
- Tüm öznitelik alt-kümesi uzayının gezilmesinden dolayı hesaplamaların çok uzun sürmesi.

### *4.3. Gömülü Çözümler (Embedded Solutions)*

Gömülü çözümlerin filtreleme çözümleri ile kapsama çözümlerinin bir birleşimi olduğu söylenebilir. Gömülü çözümler, filtreleme çözümleri gibi tüm uzayın sadece bir kısmını gezer. Kapsama çözümlerine benzer olarak ise, bir öznitelik alt-kümesini değerlendirmek için öğrenme yönteminin kalite ölçütünden faydalanır. Gömülü çözümlerin temel çalışma prensibi, öznitelik alt-kümeleri uzayı içerisinde en iyi çözümü bir arama algoritmasından faydalanarak bulmaktır [24]. Gömülü çözümlere örnek olarak LASSO, Elastic Net ve Ridge Regression verilebilir.

## 5. ÖNERİLEN ÇÖZÜM (PROPOSED SOLUTION)

Bölüm 4'te özetlenen çözümler arasından sadece filtreleme çözümleri göz önüne alınmıştır. Kapsama çözümleri ve gömülü çözümler, sınıflandırma modelinin sabit olmasını gerektirmektedir ve diferansiyel mahremiyet kullanarak birden fazla defa model oluşturulmasının anlamlı olmadığı değerlendirilmektedir. Filtreleme çözümleri için, en etkili olan değerlendiricilerin ki-kare ve bilgi kazanımı olduğu gösterilmiştir [7]. Bu sebeple bu iki değerlendiricinin diferansiyel mahremiyet ile çalışacak şekilde düzenlenmesine odaklanılmıştır.

### *5.1. Ki-kare Öznitelik Değerlendirici (Chi-squared Attribute Evaluator)*

Ki-kare değerlendirici Pearson'ın ki-kare testine dayanır. Buna göre, sınıflandırma niteliği ile seçimde değerlendirilecek öznitelik arasındaki ilişki istatistiksel olarak ölçülür. Ölçüm sonucunun ki-kare dağılımına benzer olacaktır. Bu sebeple ki-kare testi adı verilmiştir. Ki-kare testi sadece kesikli öznitelikler için uygulanabilir. Gözlemlenen dağılım ile beklenen dağılım arasındaki mesafeye dayanır ve $E_i$ beklenti, $O_i$ gözlem sayısı olmak üzere Eş. 3'deki gibi ifade edilir:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (3)$$

Örnek: Mushroom veri kümesi [25] içerisindeki renk özniteliğini ele alalım. Renklerine ve zehirli olmalarına göre veri kümesinden elde edilen kayıt sayıları Tablo 1'de verilmektedir. Bu gözlemlerden, parantez içindeki beklenti değerleri kolaylıkla hesaplanabilir. Örneğin, beyaz zehirli mantar sayısı için beklenti değer 1500*(1145/2500) = 687 olacaktır.

**Tablo 1.** Renk ve zehire göre ki-kare heabında kullanılacak gözlem ve beklentiler
(Observations and expectations to be used in chi-square calculation according to color and poisonousness)

| | Zehirli | Zehirsiz | Toplam |
|---|---|---|---|
| Beyaz | 465 (687) | 1035 (813) | 1500 |
| Kahverengi | 680 (458) | 320 (542) | 1000 |
| Toplam | 1145 | 1355 | 2500 |

Filterelemeye dayalı öznitelik seçimi yapılabilmesi için, her özniteliğe bir değer ataması yapılması gereklidir. Bu değer atamasının ki-kare ile yapılabilmesi için, her öznitelik için Tablo 1'de sunulan gözlem tablosunun oluşturulması gerekmektedir. Bu şekilde bir tablonun SQL dili ile üretilmesi oldukça kolaydır. Gruplama özelliği ile *S* sınıf niteliği, *A* ilgili öznitelik, *T* veri kümesi olmak üzere, şu sorgu yeterlidir:

```
SELECT COUNT(*), A, S
FROM T
GROUP BY A, S;
```

Bu sorgunun hassasiyeti 2 olarak hesaplanacaktır. Herhangi bir kaydın *A* veya *S* değerinin değiştirilmesi, sonucun bir hücresindeki sayıyı bir azaltacak, diğerindekini ise bir arttıracaktır. Aynı sorgunun *n* farklı öznitelik için veritabanına gönderildiği durumda ise toplam hassasiyet *2*n* olacaktır. Burada dikkat edilmesi gereken husus tek bir kayıtta yapılacak değişikliğin tüm tabloları etkilemesinin mümkün olduğudur.

### *5.2. Bilgi Kazanımı Öznitelik Değerlendirici (Information Gain Attribute Evaluator)*

Bilgi kazanımını tanıtmadan önce entropi kavramından bir miktar bahsedilecektir. Entropi, bir örneklem içerisindeki kayıtların bir nitelik üzerindeki düzensizliğini ölçmektedir [26]. Entropi kavramı bilgi teorisine dayanır ve *i* indeksli sınıf değeri $p_i$ olmak üzere *S* veri kümesinde Eş. 4'teki gibi ifade edilir:

$$\mathrm{E(S)} = -\sum_{i=1}^{n} p_i \times log_2(p_i) \quad (4)$$

Eğer *S* kümesi *A* niteliği üzerine bölümlenirse her $A_i$ olası değeri için bir $S_i$ veri kümesi oluşur. Bu bölümlemedeki entropi ise Eş. 5'teki gibi ifade edilir:

$$E_A(S) = \sum_{i=1}^{n} \frac{|S_{A_i}|}{|S|} \times E(S_i) \quad (5)$$

Yüksek entropi daha fazla düzensizlik ifade eder ve yüksek entropili nitelik üzerinde yapılacak bir bölümleme ile düzensizlik azalmış olacaktır. Bunun anlamı elde edilen bilginin artmasıdır. Buna göre bilgi kazanımı Eş. 6 ile ölçülür:

$$\mathrm{BilgiKazanımı_A} = \mathrm{E(S)} - \mathrm{E}_A\mathrm{(S)} \quad (6)$$

Bilgi kazanımı hesaplanırken *S* veri kümesinin entropisinin hesaplanmasına gerek yoktur. Her öznitelik için sadece $E_A(S)$'nin bilinmesi bilgi kazanımına göre öznitelikleri sıralamak için yeterlidir. $E_A(S)$'nin hesaplanabilmesi içinse $E(S_i)$'nin hesaplanması ve her $S_{Ai}$'nin kayıt sayısının bilinmesi yeterlidir. $E(S_i)$'yi hesaplamak için her sınıf değeri üzerinde her $A_i$ değeri için bir olasılık hesaplamak gerekir. İlgili olasılık değerleri Tablo 1'de sunulan örnekteki tablo ile kolaylıkla hesaplanabilir. Dolayısıyla, bilgi kazanımına dayalı değerlendiricinin geliştirilmesi için yine her nitelik üzerine bir gruplama sorgusu yöneltilecektir ve sorgu kümesinin hassasiyeti *2*n* olacaktır.

### *5.3. Algoritmik Karmaşıklık (Algorithmic Complexity)*

Bu bölümde, önerilen çözümlerin algoritmik karmaşıklığı analiz edilerek diferansiyel mahremiyet ile korunmuş bir öznitelik seçimi oluşturmanın getirdiği ek işlemci zamanı maliyetini ortaya konulmaktadır. Öncelikle Laplace dağılımında gürültü örneklemenin karmaşıklığı incelenecektir. Dağılımın ortalaması her zaman 0, şiddeti ise $\Delta/\varepsilon$ olarak belirlenmektedir. Bölüm 5.1 ve 5.2'de aktarılan çözümler için oluşturulan sorgu kümelerinin hassasiyeti sınıflandırmada kullanılacak öznitelik sayısı ($n$) cinsinden ifade edildiğinde sabittir ($\Delta = 2*n$) ve bu değerin hesaplanması O(1) zamanda gerçekleştirilebilir. Dolayısıyla, ilgili Laplace dağılımının belirlenmesi ve bir gürültü değerinin örneklenmesi O(1) zamanda yapılacaktır. Öznitelik seçimi yapılacak $n$ öznitelikli bir veri kümesine toplam $n$ adet gruplama sorgusu iletilmektedir ve elde edilen cevap vektörünün her hücresi bağımsız olarak gürültü ile korunmaktadır. 2 sınıf değeri ve $k$ farklı öznitelik değeri bulunduğu varsayılırsa, O($k$) zamanda gürültü ekleme işlemi tamamlanabilir. Buna göre toplam $n$ farklı öznitelik için, diferansiyel mahremiyet korumasının gerektirdiği ek maliyet sadece O($k*n$) ile ifade edilebilir. Veritabanına gönderilen gruplama sorgularının cevap vektörünün hücrelerini oluşturan ve her gruptaki kayıt sayısını içeren gürültüsüz değerleri hesaplama işleminin maliyeti ile bu değerin karşılaştırılması önemlidir. Bir grupta yer alan kayıt sayısını belirlemenin maliyeti (a) veritabanında kullanılan indeks yapılarının varlığına, (b) türüne – karma veya ağaç yapılı, (c) kümelenmiş olup olmamasına göre değişkenlik gösterecektir. Ancak en iyi senaryoda dahi (kümelenmiş, karma yapılı indeks), belirli nitelikleri sağlayan kayıt sayısının belirlenmesinin O(1)'den çok daha uzun süreceğini beklemek yerinde olacaktır. Bu sebeple, diferansiyel mahremiyet korumasının getirdiği ek hesaplama maliyeti göz ardı edilebilecek seviyede olacaktır.

### *5.4. WEKA ile Entegrasyon (Integration to WEKA)*

Yaygın olarak kullanılan bir veri madenciliği kütüphanesi olan WEKA içerisinde gözetimli ve gözetimsiz öğrenme kütüphanesi altında pek çok filtre tanımlanmıştır. Filtreler genellikle verinin ön işlemlerden geçirilmesi için kullanılır. Bu çalışmada sunulan öznitelik seçimi de bir filtre olarak Java programlama dilinde geliştirilmiş ve WEKA ile entegre çalışması sağlanmıştır. Geliştirilen bu yazılıma https://sourceforge.net/projects/dataprivacynoiseanalyser/ adresinden açık kaynak kodlu olarak erişmek mümkündür.

Bu adreste yer alan program, çalışmadaki sonuçların yeniden üretilmesini sağlayacak basit bir arayüzle donatılmıştır. Kullanıcı deney parametrelerini ilgili yazı kutularına girerek Bölüm 6'da detaylandırılacak olan tüm sonuçları doğrulayabilir. Şekil 3'te bu arayüzden bir kesit sunulmaktadır.

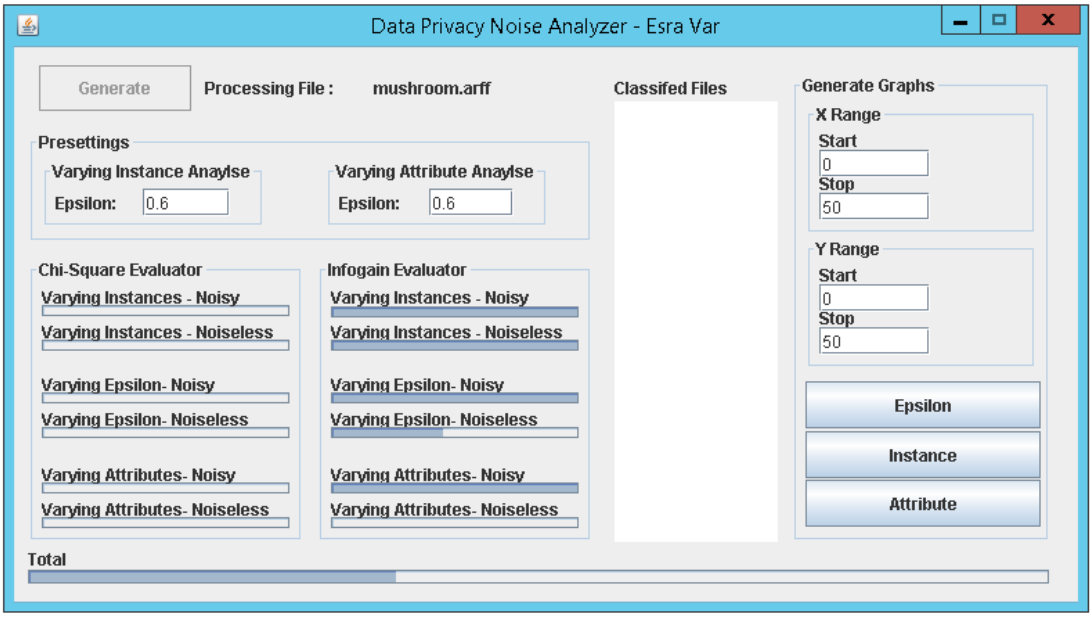
## 6. DENEY SONUÇLARI (EXPERIMENTAL RESULTS)

Bu bölümde önerilen çözümlerin gerçek veri kümeleri üzerinde elde ettiği başarım ölçülerek, kişisel gizliliği korumanın öznitelik seçimi işlemi üzerindeki etkileri vurgulanacaktır.

Başlıca deney parametreleri olarak şunlar seçilmiştir:

- Mahremiyet parametresi epsilon (ε)
- Kullanılan kayıt miktarı
- Seçilen öznitelik sayısı

Yöntemlerin kalitesini ölçmek için ise, ikili sınıflandırma yöntemleri için en geçerli ölçüt olan hata yüzdesi kullanılmıştır. Sınıf değerleri pozitif (P) ve negatif (N) olarak




**Şekil 3.** Geliştirilen çözümün kullanıcı arayüzü (User interface of the developed solution)

görülürse, hata yüzdesi hatalı bir şekilde pozitif olarak sınıflandırılmış kayıtların (hatalı pozitif - HP) ve hatalı bir şekilde negatif olarak sınıflandırılmış kayıtların (hatalı negatif - HN) boyutu |Test| olan bütün test verisi içindeki yüzdesidir. Sunulan deney sonuçlarında "hatalı sınıflandırma yüzdesi" şeklinde ifade edilen bu değerin formülü Eş. 7'de sunulmaktadır:

$$\text{Hatalı sınıflandırma yüzdesi} = 100 \times \frac{\text{HP} + \text{HN}}{|\text{Test}|} \quad (7)$$

Bütün deney sonuçları veri kümesinin 3 eşit parçaya bölünmesi ve 2 parçanın eğitim, 3. parçanın test amaçlı kullanılması yoluyla 3 defa tekrarlanmıştır. Belirtilen deney sonuçları bu 3 tekrar için ortalama değerdir. Deneylerde, öznitelik seçiminin başarımını ölçmek için, WEKA içerisinde yer alan sınıflandırma yöntemleri arasından 3 tanesi seçilmiştir:

- Naive Bayes: istatistiksel bir yöntem olması, pek çok sınıflandırma problemine kolaylıkla uygulanabilmesi ve yaygın olarak kullanılması sebebiyle seçilmiştir.
- C4.5 karar ağacı çıkartma: karar ağacının düğümleri arasında tercih yaparken bilgi kazanımından faydalandığı için seçilmiştir.
- Rastgele ormanlar: Jagannathan vd. [27] tarafından diferansiyel mahremiyete dayalı olarak geliştirildiği için seçilmiştir. Ancak bu çalışmanın amacı bütün sınıflandırma işlemini diferansiyel mahremiyete dayalı gerçekleştirmenin etkilerini incelemek olmadığı için deneylerde [27]'de önerilen gürültü sınıflandırma yönteminden faydalanılmamaktadır.

Deney sonuçları sunulurken bu sınıflandırma yöntemlerinin hiçbir öznitelik seçimi uygulanmaksızın elde ettiği başarım sırasıyla "Naive Bayes", "C4.5 Karar ağacı", ve "Rastgele ormanlar" olarak gösterilmektedir. Her deney senaryosu için bu sonuçlardan birine ek olarak 4 farklı sonuç eğrisi verilmektedir: (a) Bölüm 5.1'de sunulan ki-kare yöntemi "KiKare-DM", (b) Bölüm 5.2'de sunulan bilgi kazanımı yöntemi "BilgiKazanımı-DM" ve bunların koruma kullanılmadan uygulandığı (c) "KiKare" ve (d) "BilgiKazanımı". Her yöntem için 2 eğri sunulmasının amacı, diferansiyel mahremiyet yönteminin getirdiği kişisel gizliliği koruma niteliğine karşılık hata oranına olan etkisini görebilmektir. Önerilen diferansiyel mahremiyete dayalı çözümün, literatürde var olan diğer kişisel gizliliği koruma amaçlı yöntemler ile karşılaştırılması ne yazık ki anlamlı değildir. Şifrelemeye dayalı çözüm yöntemlerinde gizlilik koruması saldırganın şifreleme anahtarlarına sahip olmamasına dayanır. Şifreleme işlemleri analiz sonuçlarını etkilemez, sonuçlar koruma kullanılmadan elde edilen sonuçlarla birebir aynıdır. Anonimleştirmeye dayalı çözümlerde ise, veri sahibinin öznitelik seçimini uygulamasına gerek bulunmamaktadır. Veri sahibi yarı-tanımlayıcı ve hassas nitelikleri anonimlik tanımına uygun bir şekilde genelleştirecek veya bastıracaktır. Bu işlem sonucu elde edilen veri kümesi üçüncü şahıslarla olduğu gibi paylaşılabilir, amaca uygun bir şekilde – ilave gizlilik koruması gerekmeksizin – öznitelik seçimine tabi tutulabilir.

### *6.1. Veri Kümeleri (Datasets)*

Deneylerde kullanılan veriden öncelikle herhangi bir öznitelik değeri eksik olan bütün kayıtlar silinmiştir. Bu işleme "liste boyunca silme" denilmektedir [28]. Silinen kayıtların sayısı oldukça azdır, ve bu işlemin deney sonuçlarını etkilemiş olması beklenmemektedir. Liste boyunca silme işleminin özellikle sınıflandırma nitelik değeri eksik olan kayıtlar için kullanılması veri madenciliğinde standart bir uygulamadır. Tablo 2'de kullanılan veri kümelerinin temel özellikleri aktarılmaktadır. Tüm veri kümeleri UCI Makine Öğrenmesi Havuzu [25] ve Promise Havuzu [29] üzerinden açık erişim sağlanmış veri kümeleridir. Bu veri kümelerinin seçilmesinin çok temel gerekçeleri bulunmaktadır. Adult veri kümesi kişisel gizlilik çalışmalarında çok yoğun olarak kullanılması ve hem numerik, hem kategorik öznitelikler içermesi sebebiyle tercih edilmiştir. Software Defect veri kümesi öznitelik seçimi yöntemlerinin karşılaştırması için yoğun olarak kullanılması ve sadece numerik öznitelikler içermesi sebebiyle seçilmiştir. Connect-4 veri kümesi çok fazla kayıt ve öznitelik içermesi, son olarak Car veri kümesi ise çok az kayıt ve öznitelik içermesi sebebiyle tercih edilmiştir. Car ve Connect-4 veri kümelerinin tüm öznitelikleri kategoriktir. Her veri kümesi için, yukarıda belirtilen deney parametrelerinin varsayılan değerleri ise Tablo 3'de gösterildiği gibi belirlenmiştir. Epsilon veri kümesinin kayıt sayısından ve sınıflandırma probleminin zorluğundan bağımsız olarak 0,5 olarak atanmaktadır. Deneylerde veri kümesinin tamamı kullanılmaktadır ve var olan özniteliklerin yarısının seçilmesi gerekmektedir. 4 farklı veri kümesi, 3 farklı sınıflandırma yöntemi, 3 farklı deney parametresi olmak üzere toplam 4*3*3 = 36 farklı deney senaryosu bulunmaktadır. Her senaryo içinse, 5 farklı sonuç üretilmiştir. Bölüm 6.5'te farklı veri kümeleri, Bölüm 6.6'da

**Tablo 2.** Veri kümelerinin temel özellikleri (Basic properties of the datasets)

| Veri kümesi | Öznitelik sayısı | Kayıt sayısı | Sınıf sayısı | Naive Bayes hata oranı | C4.5 karar ağacı hata oranı | Rastgele ormanlar hata oranı |
|---|---|---|---|---|---|---|
| Adult | 14 | 48841 | 2 | %17,28 | %14,52 | %16,64 |
| Connect-4 | 42 | 67557 | 2 | %23,65 | %14,6 | %15,97 |
| Software Defect | 21 | 498 | 2 | %14,85 | %12,00 | %10.24 |
| Car | 6 | 1728 | 2 | %10,10 | %5,83 | %3,01 |

ise farklı sınıflandırma yöntemleri için tartışılacağı üzere deney parametrelerinin etkisi veri kümesinden ve sınıflandırma yönteminden bağımsızdır. Bu sebeple, her deney senaryosu için yalnızca en karakteristik sonuçlara yer verilmektedir.

**Tablo 3.** Veri kümeleri için varsayılan deney parametresi değerleri (Default experiment parameter values for datasets)

| Veri kümesi | Epsilon | Kayıt yüzdesi | Nitelik Sayısı |
|---|---|---|---|
| Adult | 0,5 | 100 | 7 |
| Connect-4 | 0,5 | 100 | 21 |
| Software Defect | 0,5 | 100 | 10 |
| Car | 0,5 | 100 | 3 |

### *6.2. Mahremiyet parametresi ε (Privacy Parameter ε)*

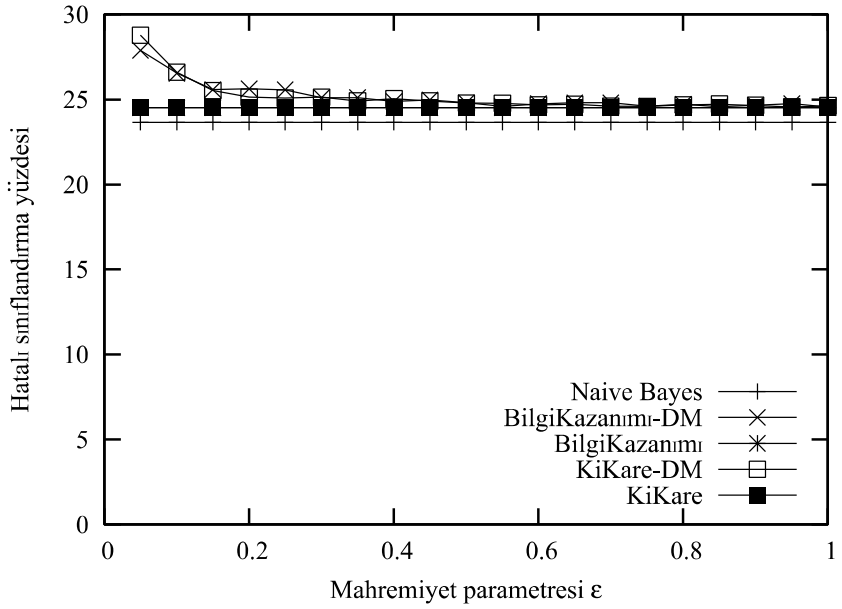
Tanım 3.2'den anlaşılacağı üzere, diferansiyel mahremiyetin başlıca koruma parametresi olan ε büyüdükçe mekanizmanın sağladığı güvenlik seviyesi düşmektedir. Buna göre daha yüksek ε, daha az gürültü cevaplar alınmasını sağlar. Bu sebeple yüksek ε değerlerinde KiKare-DM sonuçlarının KiKare sonuçlarına ve BilgiKazanımı-DM sonuçlarının BilgiKazanımı sonuçlarına yakınsaması beklenir. Şekil 4 Naive Bayes sınıflandırma yöntemi için Connect-4 veri kümesi üzerindeki sonuçları göstermektedir. ε büyüdükçe bütün öznitelik seçimli yöntemler aynı hata oranına yaklaşmaktadır. Küçük ε değerleri içinse, eklenen gürültüden dolayı KiKare-DM ve BilgiKazanımı-DM yöntemleri doğru öznitelikleri tam olarak seçememekte ve daha fazla hatalı sınıflandırma yapmaktadır.

### *6.3. Kullanılan Kayıt Miktarı (Amount of Records Used)*

Bu deney grubunda ilgili veri kümesindeki kayıtların rastgele olarak belirli bir kısmı kenara ayırılmış ve veri kümesindeki kayıt sayısının sınıflandırma hatasına etkileri incelenmiştir. Bölüm 3'te bahsedildiği üzere, diferansiyel mahremiyet tarafından sorgu cevaplarına eklenen gürültü miktarı veri kümesinin tüm niteliklerinden bağımsızdır ve sorguların bir fonksiyonuna göre (bknz. Tanım 3.1, sorgu hassasiyeti) belirlenmektedir. Kayıt miktarındaki değişiklikler, sorgu cevaplarına eklenen gürültü miktarında değişikliğe sebep olmaz. Kullanılan kayıt miktarı arttıkça, veritabanından gelen doğru cevabın şiddeti (yani, sinyal gücü) gürültü şiddetine oranla çoğalmaktadır. Doğru cevabın şiddeti arttıkça ise, KiKare-DM ve BilgiKazanımı-DM yöntemlerinin sırasıyla korumasız KiKare ve Bilgi-Kazanımı yöntemlerine yakınsamasını beklemek yerinde olur. Şekil 5 bu durumu Connect-4 veri kümesi üzerinde Naive Bayes sınıflandırma yöntemi için deneysel olarak ortaya koymakta ve doğrulamaktadır. %25 veya daha az miktarda kayıt kullanıldığı takdirde korumalı yöntemler KiKare-DM ve BilgiKazanımı-DM hatalı sınıflandırma yüzdesinde büyük oranda olumsuz farklılığa sebep olmaktadır. %25'in üzerinde ise, bu yöntemlerin hata oranı KiKare ve Bilgi-Kazanımı yöntemlerine çok yaklaşmaktadır.

### *6.4. Seçilen Öznitelik Sayısı (Number of Attributes Selected)*

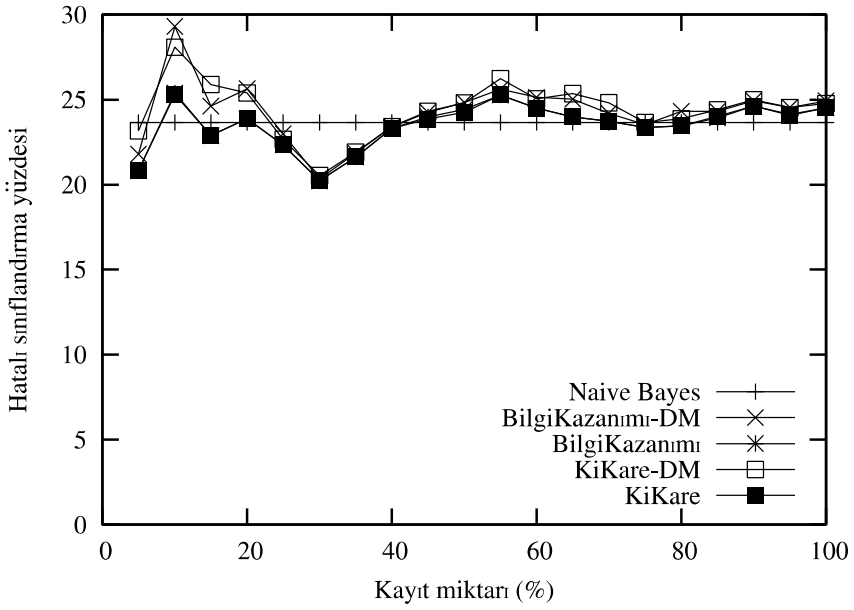
En önemli deney parametrelerinden bir tanesi olan seçilecek olan öznitelik sayısı, deney sonuçlarını iki farklı yönden etkilemektedir. Veri kümesinin $k$ öznitelik içerdiği ve bunlardan en iyi $n^{opt}$ tanesini kullanmanın sınıflandırma



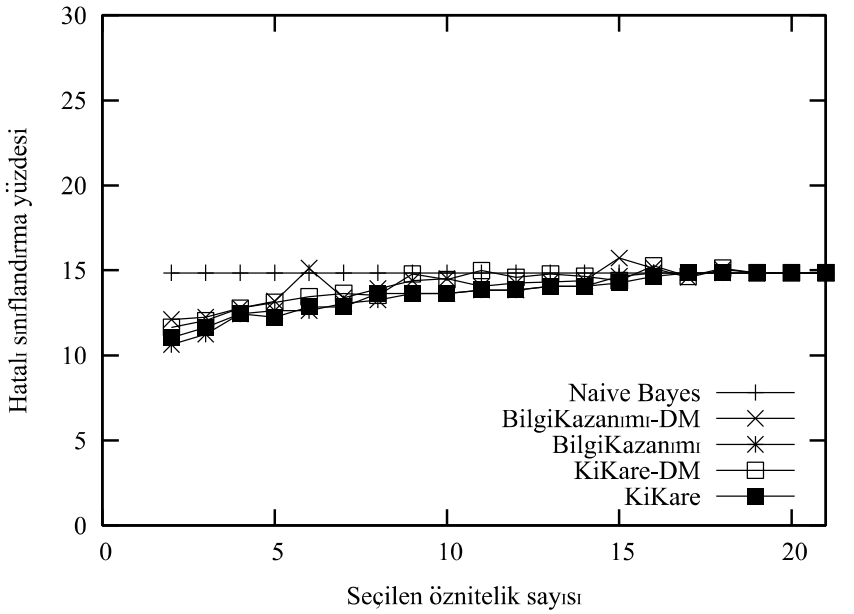

**Şekil 4.** Connect-4 veri kümesi için epsilona bağlı Naive Bayes hatalı sınıflandırma yüzdesi
(Percentage of Naive Bayes classification error on the Connect-4 dataset with varying epsilon)

işlemi için optimal bir tercih olduğu varsayılsın. Deneyde seçilen öznitelik sayısı $n$ ile ifade edilecek olursa $n$ arttıkça öznitelik seçiminin sorgu hassasiyeti olan $\Delta$ da artmaktadır. Mahremiyet bütçesi $\varepsilon$ sabit olduğu için veritabanından gelen sonuçlara daha yüksek gürültü eklenmesi gerekir.

Bu durumun ise hatalı sınıflandırma yüzdesini arttırması beklenir, çünkü öznitelik seçimi giderek daha gürültülü bir ortamda gerçekleşecek, faydalı özniteliklerin seçilmesi zorlaşacaktır. Öte yandan $n^{opt} < n \approx k$ ise, yani seçilecek öznitelik sayısı faydalı olanlardan çok daha fazla ise – hemen tüm öznitelikler seçiliyorsa – hatalı sınıflandırma yüzdesinin korumasız yöntemlere yakınsaması beklenir. Şekil 6'da bu durum Software Defect veri kümesi için Naive Bayes sınıflandırma yöntemi üzerinde gözlemlenmektedir. Bu veri kümesi için $n^{opt} = 2$ olmalıdır. $n$ büyüdükçe hatalı sınıflandırma oranı yükselmektedir. Ayrıca $n$ büyüdükçe KiKare-DM ve BilgiKazanımı-DM yöntemleri ile KiKare ve BilgiKazanımı yöntemleri arasındaki hata yüzdesi farkı yükselmektedir. Ancak, $n$ veri kümesindeki öznitelik sayısı




**Şekil 5.** Connect-4 veri kümesi için kayıt miktarına bağlı Naive Bayes hatalı sınıflandırma yüzdesi
(Percentage of Naive Bayes classification error on the Connect-4 dataset with varying amount of records)

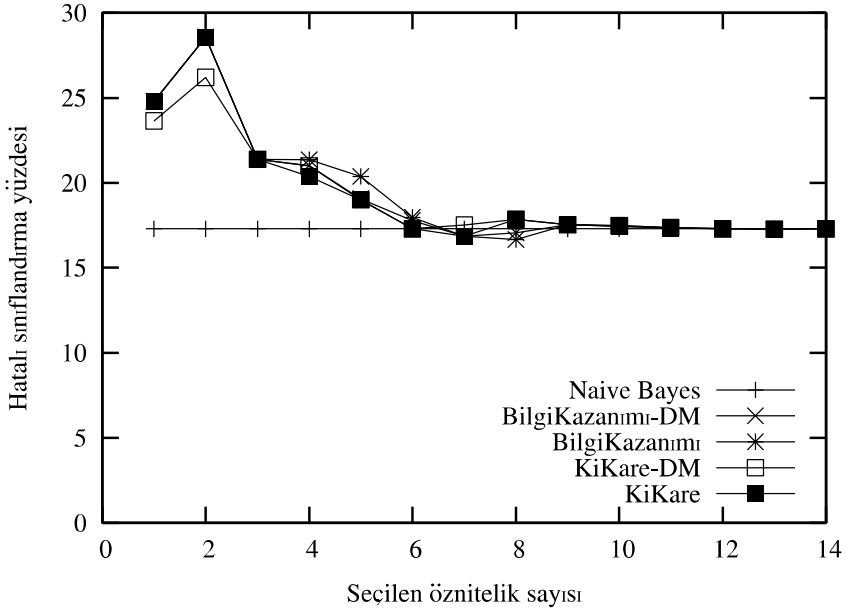



**Şekil 6.** Software Defect veri kümesi için seçilen öznitelik sayısına bağlı Naive Bayes hatalı sınıflandırma yüzdesi
(Percentage of Naive Bayes classification error on the Software Defect dataset with varying number of selected attributes)

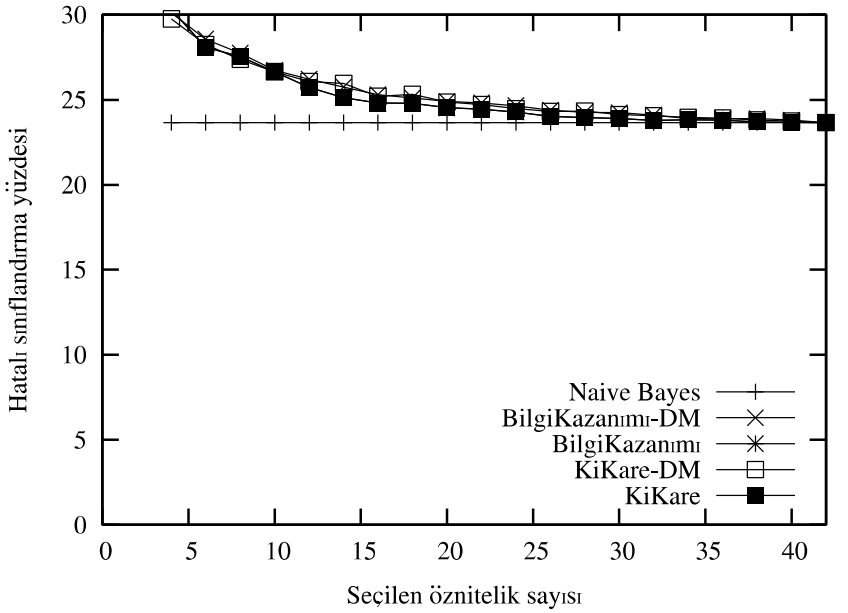
*k*'ya yaklaştıkça bütün yöntemlerin başarımı aynı noktada birleşmektedir.

### *6.5. Farklı Veri Kümeleri (Different Data Sets)*

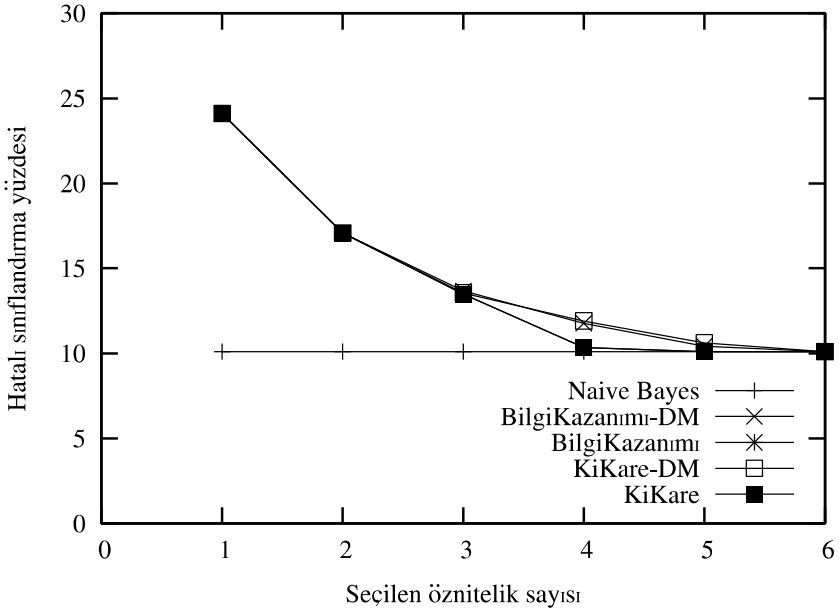
Kişisel gizliliği koruyan öznitelik seçiminin farklı veri kümeleri üzerindeki etkisini değerlendirmek amacıyla aynı koruma miktarı (ε = 0,5) ile veri kümesindeki bütün kayıtların kullanıldığı durumda (kayıt miktarı = %100) seçilen öznitelik sayısı incelenecektir. Şekil 6'da Software Defect sonuçları sunulmuştur. Şekil 7'de Adult, Şekil 8'de Connect-4 ve Şekil 9'da Car veri kümelerinin sonuçları tartışılacaktır. Adult veri kümesinde öznitelik seçimi işlemi $n < 6$ için ve $n > 9$ için sınıflandırma hatasını arttırmaktadır. Özniteliklerin bilgi kazanımı ve ki-kare ölçütlerine göre sıralaması çok fazla değişmediği için iki yöntemi birbirinden ayırmak mümkün değildir. Diferansiyel mahremiyete dayalı çözümler genellikle aynı hata miktarını vermektedir. $n = 1$ ve $n = 2$ için KiKare-DM'in beklenenden iyi sonuç verdiği gözlemlenmektedir. Bu durum incelenmiş ve sebebin




**Şekil 7.** Adult veri kümesi için seçilen öznitelik sayısına bağlı Naive Bayes hatalı sınıflandırma yüzdesi
(Percentage of Naive Bayes classification error on the Adult dataset with varying number of selected attributes)




**Şekil 8.** Connect-4 veri kümesi için seçilen öznitelik sayısına bağlı Naive Bayes hatalı sınıflandırma yüzdesi
(Percentage of Naive Bayes classification error on the Connect-4 dataset with varying number of selected attributes)




**Şekil 9.** Car veri kümesi için seçilen öznitelik sayısına bağlı Naive Bayes hatalı sınıflandırma yüzdesi
(Percentage of Naive Bayes classification error on the Car dataset with varying number of selected attributes)

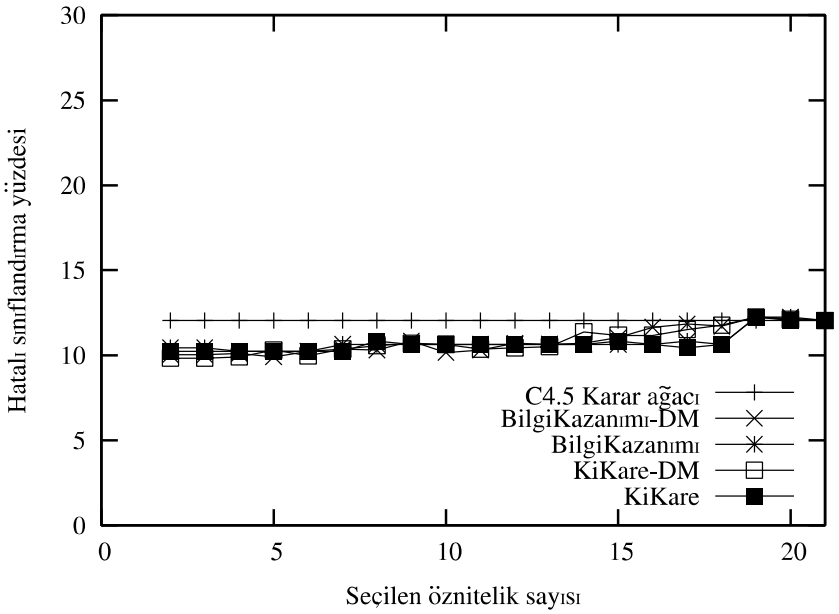
rastgele gürültünün ki-kare ölçütüne göre seçilmemesi gereken ancak sınıflandırma hatasını düşüren bir özniteliği desteklediği görülmüştür. Sebep ki-kare ölçütünün hatalı olarak elediği faydalı özniteliğin gürültü sayesinde seçilmesidir. Connect-4 veri kümesi çok fazla kayıt ve öznitelik içermektedir. Seçilen öznitelik sayısı azken, gürültü miktarı düşük kalmakta ve kişisel mahremiyet korumalı çözümler ile korumasız çözümler hemen hemen aynı hatalı sınıflandırma yüzdesine erişmektedir. Büyük *n* değerleri içinse, BilgiKazanımı ve KiKare eğrilerinden görülebileceği üzere öznitelik seçiminin sınıflandırma başarımı üzerinde herhangi bir etkisi bulunmamaktadır. Bu sebeple bütün yöntemler hızla Naive Bayes eğrisine yakınsamaktadır. Car veri kümesi kayıt sayısına göre çok az sayıda öznitelik içerdiği için, gürültünün doğru cevaplar üzerindeki etkisi çok sınırlı miktarda olmaktadır. Bu sebeple tüm hatalı sınıflandırma yüzdeleri yakın ölçülmektedir. Bölüm 6.4'te tartışıldığı üzere, *n* arttıkça (bakınız *n = 4* ve *n = 5*) korumalı yöntemlerin başarımının düştüğü bu veri kümesinde de gözlemlenebilir. *n= 6* için zaten bir seçim söz konusu değildir. Farklı veri kümeleri için elde edilen deney sonuçları birbirinden farklı olmakla birlikte, bu farklılıklar çoğunlukla veri kümesinin kayıt sayısı, öznitelik sayısı gibi öznitelik seçimini etkileyebilecek karakteristik özelliklerinden kaynaklanmaktadır. Diferansiyel mahremiyet koruması doğru sonuçlara eklenen gürültüden dolayı öznitelik seçiminde bir miktar başarısız sergilemektedir, ancak bu seçimlerin sınıflandırma başarımı üzerindeki etkileri de kısıtlıdır.

### *6.6. Farklı Sınıflandırma Yöntemleri*
*(Different Classification Methods)*

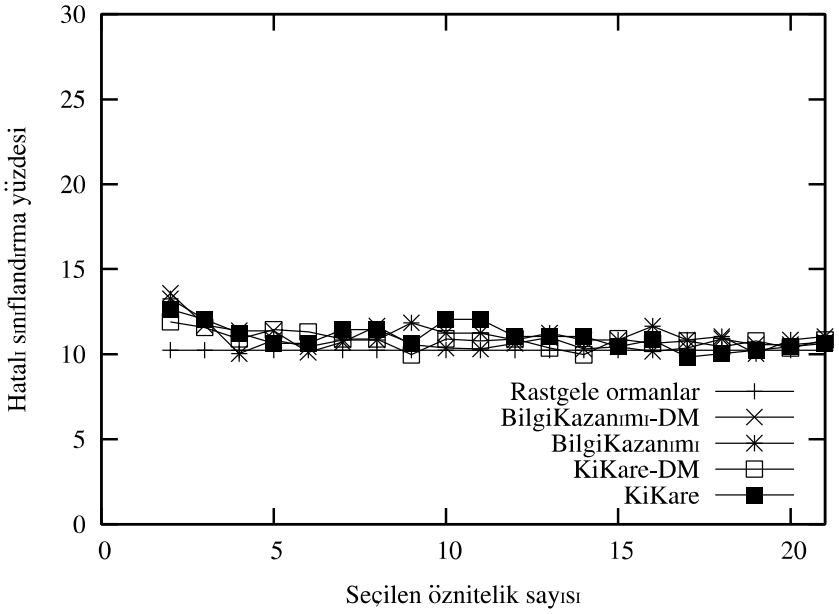
Bu aşamaya kadar sunulan tüm deney sonuçları Naive Bayes sınıflandırma yöntemini kullanmıştır. Bu bölümde farklı sınıflandırma yöntemleri ile elde edilen sonuçlar Şekil 6 ile karşılaştırmalı olarak tartışılacaktır. Sunulan tüm sonuçlarda $\varepsilon = 0{,}5$ koruma ile veri kümesindeki tüm kayıtlar kullanılmıştır. C4.5 karar ağacı çıkartma yönteminin sonuçları Şekil 10'da sunulmaktadır. *n = 13*'e kadar yöntemlerin başarımları arasında kayda değer bir farklılık görülmemektedir. Ancak daha büyük *n* değerleri için Naive Bayes yönteminde de olduğu gibi diferansiyel mahremiyet koruması sağlayan BilgiKazınımı-DM ve KiKare-DM yöntemleri daha fazla hatalı sınıflandırma yapmaktadır. *n > 18* durumunda ise Şekil 6 için görüldüğü gibi öznitelik seçimi etkisini kaybetmektedir. Rastgele ormanlar sınıflandırma modeli kendi içinde rastsallaştırma barındıran bir yöntemdir. Buna göre, seçilen özniteliklerin bir alt kümesi rastgele seçilerek bunların her biri için derinliği 1 olan bir karar ağacı oluşturulur. Sınıflandırma ise, bu karar ağaçlarından elde edilecek sonuçların bir birleşimidir (ör. oy çokluğu). Yöntemin kendi içinde rastsallık içermesi sebebiyle öznitelik seçimine dayalı bütün eğrilerin salınım halinde olduğu gözlemlenmektedir. Bu durum, bu çalışmada önerilen seçim yöntemlerinin avantajını ifade etmektedir. Zira diferansiyel mahremiyete dayalı olarak öznitelik seçiminin akabinde gerçekleşmesi gereken diferansiyel mahremiyete dayalı sınıflandırma modeli oluşturulması aşamasında, eklenen gürültünün bir kısmının iptali söz konusu olabilecektir. Bu durumda, Şekil 11'de görüldüğü üzere, öznitelik seçimi kullanımı kayda değer bir sınıflandırma hatasına sebep olmayabilir.

### *6.7. Hesaplama İşlemi Süresi (Computation Time)*

Önerilen çözümlerin sebep olduğu ekstra hesaplama maliyetleri Bölüm 5.3'te algoritmik olarak incelenmiştir. Bu bölümde sadece tek bir deney senaryosu ile ölçülen gerçek değerlerin ne kadar düşük olduğu ampirik olarak ortaya



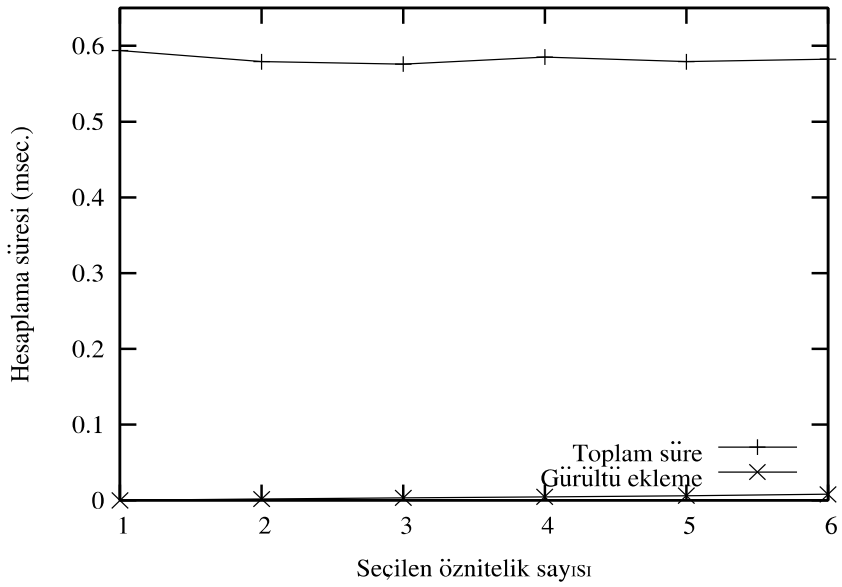

**Şekil 10.** Software Defect veri kümesi için seçilen öznitelik sayısına bağlı C4.5 karar ağacı hatalı sınıflandırma yüzdesi
Percentage of C4.5 decision tree classification error on the Software Defect dataset with varying number of selected attributes)




**Şekil 11.** Software Defect veri kümesi için seçilen öznitelik sayısına bağlı rastgele ormanlar hatalı sınıflandırma yüzdesi
(Percentage of random forests classification error on the Software Defect dataset with varying number of selected attributes)

konulmaktadır. Mahremiyet parametresi ε'un işlem süresini etkilemeyeceği açıktır. Seçilen öznitelik sayısının Bölüm 5.3'te tartışıldığı üzere lineer etkiye sahip olması beklenir. Kayıt sayısı ise seçim yönteminin çalışma süresinde etkili olmakla birlikte, diferansiyel mahremiyet koruması sağlanması için etkisizdir. BilgiKazanımı-DM ve KiKare-DM yöntemleri işlem süresi açısından farklılık göstermeyecektir – aynı şiddette gürültü aynı sayıda doğru değere eklenmektedir.

Bu sebeple Şekil 12'de verilen deney sonuçlarında seçilen öznitelik sayısına bağlı olarak sadece KiKare-DM'e ait işlem sürelerine yer verilmiştir. Ölçülen işlemci zamanı sadece öznitelik seçimi işlemine aittir. Sınıflandırma modeli kurulması, bu modelin hatalı sınıflandırma yüzdesinin tespit edilmesi gibi kısımlar ölçümün dışında tutulmuştur. Sonuçlar Bölüm 5.3'teki analizi desteklemektedir. Car veri kümesi yerine kayıt sayısı itibariyle daha büyük bir veri kümesi kullanılması halinde KiKare-DM'in işlem süresi




**Şekil 12.** Diferansiyel mahremiyet korumasının seçilen öznitelik sayısına bağlı ek hesaplama maliyeti
(Added computational cost of differential privacy protection with varying number of selected attributes)

değişmeyecek, ancak toplam işlem süresi çoğalacaktır. Dolayısıyla sunulan sonuçlar diferansiyel mahremiyet için harcanan işlem süresinin görece en yüksek olduğu duruma aittir. Ölçümler harcanan sürenin göz ardı edilebilir seviyede olduğunu doğrulamaktadır.

## 7. SONUÇLAR (CONCLUSIONS)

Filtreleme yöntemleri üzerinde etkili çalıştığı bilinen ki-kare ve bilgi kazanımına dayalı değerlendiriciler ile diferansiyel mahremiyete dayalı öznitelik seçimi çözümleri önerilmiştir. Değerlendiricilerin veritabanı erişimi sorgular şeklinde modellenmiş ve bu sorguların hassasiyetleri hesaplanarak, doğru sonuçlara diferansiyel mahremiyet uyarınca gürültü eklenmiştir.

Farklı sınıflandırma yöntemleri kullanılarak elde edilen deney sonuçlarına göre, gürültülü sonuçlara ait sınıflandırma doğruluğu, gürültüsüz sonuçların doğruluğuna oldukça yakındır. Buna göre, önerilen çözüm, doğruluktan anlamlı bir sapmaya sebep olmaksızın diferansiyel mahremiyet koruması sağlamaktadır. İleriki çalışmalarda bu değerlendiriciler için alternatif sorgulama yöntemlerinin irdelenmesi planlanmaktadır. Bir diğer olası çalışma alanı ise, sınıflandırmanın kesikleme gibi diğer ön işlemlerinin diferansiyel mahremiyete dayalı yapılmasıdır.

## KAYNAKLAR (REFERENCES)


1. Kantarcioglu M., Privacy-Preserving Distributed Data Mining And Processing On Horizontally Partitioned Data, PhD thesis, Purdue University, 08-2005.
2. Vaidya J., Privacy Preserving Data Mining over Vertically Partitioned Data, PhD thesis, Purdue University, 08-2004.
3. Sweeney L., Achieving k-anonymity privacy protection using generalization and suppression, Int. J. Uncertain. Fuzziness Knowl.-Based Syst., 10 (5), 571-588, 2002.
4. Machanavajjhala A., Kifer D., Gehrke J., Venkitasubramaniam M., l-diversity: privacy beyond k-anonymity, ACM Trans. Knowl. Discov. Data, 1 (1), 1-36, 2007.
5. Li N., Li T., t-closeness: privacy beyond k-anonymity and l-diversity, Proc. of IEEE 23rd Int'l Conf. on Data Engineering, İstanbul-Turkey, 106-115, 2007.
6. Dwork C., Differential privacy: A survey of results, Proc. of the 5th International Conference on Theory and Applications of Models of Computation, Heidelberg-Berlin, 1-19, 2008.
7. Yang Y., Pedersen J.O., A comparative study on feature selection in text categorization, Proc. of the Fourteenth International Conference on Machine Learning, San Francisco CA - USA, 412-420, 1997.
8. Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. H., The weka data mining software: An update. SIGKDD Explor. Newsl., 11 (1), 10-18, 2009.
9. Aggarwal C.C., On k-anonymity and the curse of dimensionality, Proc. of the 31st International Conference on Very Large Data Bases, Trondheim-Norway, 901-909, 2005.
10. Zhang M.M., ZouG L., A new data perturbation method of reference control in statistical database. Applied Mechanics and Materials, 241, 3134-3137, Trans. Tech. Publications, 2013.
11. Zayatz L., Evans T., Slanta J., Using noise for disclosure limitation of establishment tabular data, Journal of Official Statistics, 14 (4), 537-551, 1998.
12. Demirelli Okkalıoğlu B., Koç M., Polat H., Deriving private data in partitioned data-based privacy-preserving collaborative filtering systems, Journal of the

Faculty of Engineering and Architecture of Gazi University, 32 (1), 53-64, 2017.

**13.** Shlomo N., Skinner C.J., Privacy protection from sampling and perturbation in survey microdata. Journal of Privacy and Confidentiality, 4 (1), 155-169, 2012.

**14.** Kadampur M.A., Somayajulu D.V.L.N., A noise addition scheme in decision tree for privacy preserving data mining. The Computing Research Repository, arXiv:1001.3504, 2010.

**15.** Soria-Comas J., Domingo-Ferrer J., Optimal data-independent noise for differential privacy, Information Sciences, 250 (0), 200-214, 2013.

**16.** Lee D.G.Y., Protecting Patient Data Confidentiality Using Differential Privacy, MSc. Thesis, Oregon Health and Science University, 2008.

**17.** Lee N.Y., Kwon O., A privacy-aware feature selection method for solving the personalization-privacy paradox in mobile wellness healthcare services. Expert Syst. Appl., 42 (5), 2764-2771, 2015.

**18.** Gkoulalas-Divanis A., Loukides G., Sun J., Publishing data from electronic health records while preserving privacy: A survey of algorithms, Journal of Biomedical Informatics, 50, 4-19, 2014.

**19.** Çelik C., Bilge H.Ş., Feature selection with weighted conditional mutual information, Journal of the Faculty of Engineering and Architecture of Gazi University, 30 (4), 585-596, 2015.

**20.** Akben S.B., Alkan A., Density-based feature extraction to improve the classification performance in the datasets having low correlation between attributes, Journal of the Faculty of Engineering and Architecture of Gazi University, 30 (4), 597-603, 2015.

**21.** Xiao X., Tao Y., Output perturbation with query relaxation. Proc. VLDB Endow., 1 (1), 857-869, 2008.

**22.** Dwork C., McSherry F., Nissim K., Smith A., Calibrating noise to sensitivity in private data analysis, Lecture Notes in Computer Science, 3876, 265-284. Springer, Berlin Heidelberg, 2006.

**23.** John G.H., Kohavi R., Pfleger K., Irrelevant features and the subset selection problem. Proc. of the Eleventh International Conference on Machine Learning, New Brunswick NJ – USA, 121-129, 1994.

**24.** Xiao Z., Dell E., Dou W., Chen L., ESFS: A new embedded feature selection method based on SFS, Rapports de recherché, RR-LIRIS-2008-018, 1-10, 2008.

**25.** Lichman M., UCI machine learning repository, http://archive.ics.uci.edu/ml, published: 2013, accessed: Jan. 2018.

**26.** Mitchell T.M., Machine Learning, McGraw-Hill Inc., New York NY-USA, 1st edition, ISBN 0070428077, 1997.

**27.** Jagannathan G., Pillaipakkamnatt K., Wright R.N., A practical differentially private random decision tree classifier. Trans. Data Privacy, 5 (1), 273-295, 2012.

**28.** Allison P.D., Missing Data, SAGE Publications, ISBN 9780761916727, 2002.

**29.** Sayyad Shirabad J., Menzies T.J., The PROMISE Repository of Software Engineering Databases. School of Information Technology and Engineering, University of Ottawa, Canada. http://promise.site.uottawa.ca / SERepository, published: 2005, accessed: Jan. 2018.