**KİŞİSEL VERİ VE HASSASİYETİ ÜZERİNE BİRLİKTELİK ANALİZİ**

Barış Yıldız

Dokuz Eylül Üniversitesi, Türkiye

baris.yildiz@ceng.deu.edu.tr

ORCID: 0000-0003-4636-9957

Alp Kut

Dokuz Eylül Üniversitesi, Türkiye

alp@cs.deu.edu.tr

ORCID: **0000-0002-5781-334X**

**ÖZET**

Muazzam miktarda veriyi işleme yeteneğinin artmasıyla, daha önce endişelenmediğimiz sorunlarla karşılaşıyoruz: gizlilik ve hassas bilgilerin işlenmesi. Bu nedenle, veri analizi için geliştirilen sistemler gizlilik endişelerine uygun olmalıdır. Tasarım yoluyla ve erişim kontrol mekanizmaları aracılığıyla bu tür sistemler gizlilik bilinciyle geliştirilebilir. Bu da tarafların karşılıklı güvenini arttırarak sürdürülebilir bir toplumun oluşmasına katkı sağlayacaktır. Bununla ilgili olarak, kişisel veri ve hassasiyeti konusunda insanların farkındalıklarını bulmak için bir anket hazırlanmıştır. Sorular altı gruba ayrılabilir. Bunlar: a) Özel olmayan kişisel bilgiler, b) Yanıltıcı bilgi verme eğilimi, c) Farklı durumlarda mahremiyet endişelerindeki değişiklik, d) Sadakat kartları ve sağlanan bilgiler, e) Şirketler arası bilgi paylaşımı endişeleri, f) Sosyal ağlar ve konum bilgisi verme eğilimi hakkında bilgi toplamaya yöneliktir. Veri toplama sürecinde toplam 171 kayıt elde edilmiştir. Bunun ardından, birliktelik analizi için uygun bir veri kümesi oluşturulmuş ve analiz yapılmıştır. Sonuçlar, katılımcıların mümkün olduğunda mahremiyetlerini koruma eğiliminde olduklarını ve farklı durumlarda farklı bir duyarlılık tutumuna sahip olduklarını göstermektedir. Ayrıca katılımcıların kendilerini rahat hissetmediklerinde yanıltıcı bilgi verme eğiliminde olduğu gözlemlenmiştir. Bu çalışma, kişilerin hassas bilgilerin toplanmasından çok üçüncü kişilerle paylaşılmasından rahatsızlık duyduklarını göstermektedir.

**Anahtar Kelimeler:** kişisel veri, gizlilik, veri madenciliği

**ASSOCIATION ANALYSIS ON PERSONAL DATA AND ITS SENSITIVITY**

Barış Yıldız

Dokuz Eylul University, Turkey

baris.yildiz@ceng.deu.edu.tr

ORCID: 0000-0003-4636-9957

Alp Kut

Dokuz Eylul University, Turkey

alp@cs.deu.edu.tr

ORCID: **0000-0002-5781-334X**

**ABSTRACT**

With the growing ability to process enormous volumes of data, we come across problems that we did not worry about earlier: privacy and processing of sensitive information. Therefore, developed systems for data analysis should comply with privacy concerns. By design and through access control mechanisms such systems can be developed in awareness of privacy. This will increase the mutual trust of the parties and contribute to the formation of a sustainable society. In this regard, a questionnaire was prepared to find out people's awareness about personal data and sensitivity. With this questionnaire of six group of questions, information about: a) Non-private personal information, b) Tendency to provide misleading information, c) Change in privacy concerns in different situations, d) Loyalty cards and information provided, e) Concerns about sharing information between companies, f) Social networks and information about the tendency to give location information are collected. A total of 171 records were obtained during the data collection process. After that, a suitable dataset for association analysis was created and analysis was made. The results show that participants tend to protect their privacy whenever possible and have a different attitude of sensitivity in different situations. In addition, it has been observed that participants tend to give misleading information when they do not feel comfortable. This study shows that people are uncomfortable with sharing sensitive information with third parties rather than collecting it.

**Keywords:** personal data, privacy, data mining

1. **Giriş**

Veri üretim hızındaki artış, veri kaydetme maliyetlerinin azalması ve veri işleme kabiliyetlerinin artması ile birlikte geçmişte analiz edilemeyen ya da analiz etmeye zahmet edilmeyen veri kümeleri de artık analiz edilir olmuştur. Dolayısı ile hassas nitelikli veri de işlenmektedir. Ek olarak bazı durumlarda verinin kendisi değil işlendikten sonra çıkan sonuç da hassas nitelikli olabilmektedir. Bunun yanında bazı çıkartım mekanizmaları yoluyla hassas nitelikli analiz sonuçları ilişkilendirilerek mahremiyet sorunu doğurabilmektedir. Bütün bunların bir sonucu olarak da günümüzde hassas verilerin toplanması ve işlenmesi konusu kamuoyunun gündeminde olan bir konu halini almıştır. Amerika(*Health Information Privacy*, 2015), Avrupa(*General Data Protection Regulation (GDPR) – Official Legal Text*, n.d.) ve Türkiye’de(*KİŞİSEL VERİLERİ KORUMA KURUMU | KVKK | Kişisel Verileri Koruma Kurumu Başkanlığı*, n.d.) de çeşitli yasalarla kişisel verilerin korunmasına yönelik çalışmalar yapılmaktadır. Hasta kayıt sistemleri, sosyal ağlar, akıllı ev/şehir ve sanal alışveriş gibi birçok ortamda kişisel hassas verileri de içeren çok fazla veri üretilmektedir.

Büyük miktarda veriden değerli bilgileri bulmak veri madenciliği olarak tanımlanabilir(Liu & Özsu, 2009). Temel olarak üç veri madenciliği tekniği vardır: sınıflandırma, kümeleme ve birliktelik kuralı madenciliği(Han & Kamber, 2005). Sınıflandırma, bir eğitim seti kullanır ve yeni örneklerin sınıflarını tahmin etmek için bir sınıflandırıcı oluşturur. Kümeleme, veri kümesini, üyeleri birbirine benzer ve diğer kümelerin üyelerinden farklı olan kümelere böler. Birliktelik kuralı madenciliği, veri kümesi arasındaki kalıpları ve ilişkileri bulur. Birliktelik kuralı madenciliği, sık öğe kümesi veya sık örüntü madenciliğini içerir. Biz bu çalışmada hazırlanan anket sonuçları üzerinde birliktelik analizi yaparak sık karşılaşılan örüntüleri bulmaya çalıştık. Takip eden bölümlerde hazırlanan anket ve yapılan analiz ile ilgili bilgiyi bulabilirsiniz.

1. **Kişisel Veri ve Hassasiyeti Farkındalık Anketi**

Kişisel veri ve hassasiyeti konusunda insanların farkındalıklarını bulmak için bir anket hazırlanmıştır. Bu anket 25 sorudan oluşmaktadır. Bu sorular edinilmek istenen bilgiye göre 6 gruba ayrılabilir.

* Özel olmayan kişisel bilgiler: Yaş, cinsiyet ve eğitim durumu bilgilerinin alındığı sorulardır.
* Yanıltıcı bilgi verme eğilimi: Formlar ve anketler gibi bilgi girişi olan durumlarda katılımcıların eksik ya da yanlış bilgi verme eğilimlerinin sorgulandığı sorulardır.
* Farklı durumlarda mahremiyet endişelerindeki değişiklik: Toplumda baskı konusu olabilecek durumlarda kişinin mahremiyet algısındaki değişikliğin sorgulandığı sorulardır. Hastanede farklı polikliniklerde bulunma senaryosu üzerinden sorular yansıtılmıştır.
* Sadakat kartları ve sağlanan bilgiler: Genelde indirim kazanmak için kullanılan sadakat kartlarının üyeliklerinde üye olunan firmaya verilen bilgilerin sorgulandığı sorulardır.
* Şirketler arası bilgi paylaşımı endişeleri: Bilgi sağlamadığı kendisi ile ilgili bilgi elde etmiş firmaların katılımcılara ulaşma şekilleri ve ne derece rahatsızlık duyduklarının sorgulandığı sorulardır.
* Sosyal ağlar ve konum bilgisi verme eğilimi: Sosyal ağları kullanma durumları ve bilgi paylaşımında nelere dikkat ettiklerinin sorgulandığı sorulardır.

Çevrimiçi ortamda hazırlanan anket ile 13.08.2020 – 30.10.2020 tarihleri arasında katılımcıların kayıtları alındı. Bu süreçte 171 kişiye ulaşıldı. Katılımcıların yaş ve cinsiyet dağılımları aşağıdaki grafikte de görülebileceği gibi dengeli olmuştur. Bu açıdan bakıldığında elde ettiğimiz veri kümemizin oldukça değerli olduğunu dile getirilebiliriz.

Şekil . Katılımcıların yaş ve cinsiyet dağılımı



1. **Birliktelik Analizi**

İlk olarak (Agrawal et al., 1993)'de bahsedilen birliktelik kuralı madenciliği, en popüler veri madenciliği yaklaşımlarından biridir. Market sepeti verilerinde başka bir ürün satın alınırken bir ürünün de satın alınıyor olması bir birlikteliği temsil eder. En meşhur ve açıklayıcı örneği “Bebek Bezi - Bira”’dır. Amerika’da babaların bebeklerine bebek bezi aldıklarında, hafta sonu gerçekleşecek maçları izlemek için aynı zamanda bira da almaları ile açıklanabilir(Dunham, 2002). Benzer durumau günlük yaşamımızda market raflarında ürün yerleştirme ve kampanyalı ürünler olarak oldukça sık görmekteyiz. Bazı popüler birliktelik kuralı madenciliği algoritmaları şunlardır: Apriori(Agrawal & Srikant, 1994), ECLAT (Zaki et al., 1997) ve FP-Growth(Han et al., 2000).

Örnek olarak aşağıdaki tablodaki her bir satırın bir alışverişi temsil ettiğini düşünün. Bu durumda 10 alışverişlik bir veri kümemiz bulunmaktadır. 5 çeşit de ürünümüz bulunmaktadır.

Tablo 1. Örnek veri kümesi



Bu örnek kümedeki minimum sıklık(destek) değerini 3(%30) seçersek en az 3 satırda birlikte geçen elemanları, market örneği üzerinden düşünülürse de en az 3 alışverişte birlikte alınan ürünleri bulmaya çalışıyoruz demektir. Bu durumda {A},{C},{D},{E},{AC},{CD} kümeleri sık geçen kümeler olarak bulunur.

Anket verisi üzerinde birliktelik analizi yapmak için öncelikle anket sonuçlarını uygun biçime sokmak gerekmektedir. Bu durumda bir sorunun işaretlenen şıkkı bizim için bir elemanı temsil etmektedir. 25 sorudan oluşan anketimizde bu bize 113 elemanlı bir kümeyi vermektedir. 2^113 olası kümenin 171 katılımcıdan oluşan anket sonuçlarında kaçında bir arada geçtiğinin bulunması gerekmektedir. Neyse ki daha önce sözü edilen algoritmalar sayesinde sık geçen kümeler kolaylıkla bulunabilmektedir. Biz çalışmamızda (Borgelt, 2018)’de verilen Apriori algoritmasını kullandık.

Aşağıdaki tabloda en sık geçen tek elemanlı olmayan 5 birliktelik ve bu birlikteliklerin destek değerleri verilmiştir.

Tablo 2. En yüksek destek değerine sahip ilk beş birliktelik



 Basit istatiksel sonuçlardan farklı olarak bizim için en ilginç birliktelikler 11 ve 13 numaralı sorulara verilen yanıtlardan çıkartılan birliktelikler olmuştur. Bu örüntüleri şu şekilde yorumlayabiliriz: katılımcılar kişisel verilerini paylaşmaktadır ancak bu verilerin üçüncü kişiler tarafından kullanılmasından oldukça fazla rahatsızlık duymaktadırlar.

1. **Sonuç**

Sonuç olarak çalışmamız göstermiştir ki katılımcılar mümkün olduğunda mahremiyetlerini koruma eğilimindedirler. Farklı durumlarda farklı bir duyarlılık tutumuna sahip olduklarını göstermektedir. Ayrıca katılımcıların kendilerini rahat hissetmediklerinde yanıltıcı bilgi verme eğiliminde olduğu gözlemlenmiştir. Çalışmanın en önemli sonucu ise kişilerin kendileri ile ilgili veri toplanması ve işlenmesinden çok üçüncü kişilerle olan paylaşımdan rahatsızlık duyduklarıdır. Bu durumun toplumun sürdürülebilirliği için sorun teşkil ettiği söylenebilir. Çeşitli mekanizmaların kullanılarak kişilerin mahremiyetinin koruma altına alınması ve karşılıklı güvenin sağlanması gerekmektedir.

**KAYNAKÇA**

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, *22*(2), 207–216. https://doi.org/10.1145/170036.170072

Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. *VLDB ’94 Proceedings of the 20th International Conference on Very Large Data Bases*, 487–499. http://www.vldb.org/conf/1994/P487.PDF

Borgelt, C. (2018). *Christian Borglet’s Web Page*. http://www.borgelt.net/eclat.html

Dunham, M. H. (2002). *Data Mining : Introductory and Advanced Topics*. Prentice Hall. https://www.pearson.com/us/higher-education/program/Dunham-Data-Mining-Introductory-and-Advanced-Topics/PGM324454.html

*General Data Protection Regulation (GDPR) – Official Legal Text*. (n.d.). Retrieved October 21, 2020, from https://gdpr-info.eu/

Han, J., & Kamber, M. (2005). *Data Mining : Concepts and Techniques* (2nd ed., Issue March). Morgan Kaufmann.

Han, J., Pei, J., Yin, Y., Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data - SIGMOD ’00*, *29*(2), 1–12. https://doi.org/10.1145/342009.335372

*Health Information Privacy*. (2015, August). https://www.hhs.gov/hipaa/index.html

*KİŞİSEL VERİLERİ KORUMA KURUMU | KVKK | Kişisel Verileri Koruma Kurumu Başkanlığı*. (n.d.). Retrieved October 21, 2020, from https://www.kvkk.gov.tr/

Liu, L., & Özsu, M. T. (2009). *Encyclopedia of database systems*. Springer.

Zaki, M. J., Parthasarathy, S., Ogihara, M., & Li, W. (1997). New algorithms for fast discovery of association rules. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining* (pp. 283–286). AAAI Press. http://dl.acm.org/citation.cfm?id=3001454