T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ARAŞTIRMA PROJELERİNDE KÜMELEME İLE ÇOKLU ANALİZ

YÜKSEK LİSANS TEZİ

Gülizar PAT

Enstitü Anabilim Dalı	**:**	**BİLİŞİM SİSTEMLERİ MÜHENDİSLİĞİ**
Tez Danışmanı	**:**	**Prof. Dr. Ümit KOCABIÇAK**

Temmuz 2020

BEYAN

Tez içindeki tüm verilerin akademik kurallar çerçevesinde tarafımdan elde edildiğini, görsel ve yazılı tüm bilgi ve sonuçların akademik ve etik kurallara uygun şekilde sunulduğunu, kullanılan verilerde herhangi bir tahrifat yapılmadığını, başkalarının eserlerinden yararlanılması durumunda bilimsel normlara uygun olarak atıfta bulunulduğunu, tezde yer alan verilerin bu üniversite veya başka bir üniversitede herhangi bir tez çalışmasında kullanılmadığını beyan ederim.

Gülizar PAT

20.08.2020

TEŞEKKÜR

Yüksek lisans eğitimim boyunca değerli bilgi ve deneyimlerinden yararlandığım, her konuda bilgi ve desteğini almaktan çekinmediğim, araştırmanın planlanmasından yazılmasına kadar tüm aşamalarında yardımlarını esirgemeyen, teşvik eden, aynı titizlikte beni yönlendiren değerli danışman hocam Prof. Dr. Ümit KOCABIÇAK'a teşekkürlerimi sunarım.

Ayrıca bu çalışmanın maddi açıdan desteklenmesine olanak sağlayan Sakarya Üniversitesi Bilimsel Araştırma Projeleri (BAP) Komisyon Başkanlığına (Proje No: 2020-7-24-45) teşekkür ederim.

İÇİNDEKİLER

SİMGELER VE KISALTMALAR LİSTESİ

YÖK	: Yüksek Öğretim Kurumu
EM	: Expectation Maximization
SABİS	: Sakarya Üniversitesi Bilgi Sistemi
SOM	: Self-Organizing Maps

ŞEKİLLER LİSTESİ

TABLOLAR LİSTESİ

ÖZET

Anahtar kelimeler: Veri Madenciliği, Kümeleme Analizi, Araştırma Projeleri

Bu çalışmada, Sakarya Üniversitesi bilgi sisteminde yer alan bilimsel araştırma proje yönetim sistemi üzerinden gerçekleştirilen araştırma projeleri için oluşturulan başvuru taslakları ve gerçekleştirilen başvuruları bu başvuruların değerlendirilme süreçleri incelenmiş ve proje çalışma alanlarına göre kümeleme ile çoklu analiz gerçekleştirilmiştir.

Daha önce yapılan araştırma projelerinin başvurularının kabul süreci gerçekleşmeden hatta değerlendirme işleminden sonra red alan projelere ait sağlıklı bir veri kaydı bulunmamakta idi. 2018 yılında sanal ortamda proje oluşturma ve başvuru safhalarına geçiş ile artık proje taslakları ve araştırma projelerine ait başvuruların tüm safhaları kayıt altına alınmaktadır. Proje başvurusu esnasında proje yürütücüsü tarafından anahtar kelimelerin kaydı yapılırken ilgili temel alan ve bilim alanları bilgileri istenmektedir. Başvuruların çalışma alanlarına ait kurumun verdiği destek ve gelecek yatırım alanlarını belirleyen model gösterilmiştir.

Kümeleme analizi ile araştırma projelerinin çalışma alanları projelere ait anahtar kelimeler üzerinden belirlenerek veri madenciliği tekniklerinden kümeleme analizi için K-Means, X-Means ve EM (Expectation Maximization) algoritmaları uygulanarak ilgili algoritmaların aralarındaki değerlendirme yapılmıştır.

MULTIPLE ANALYSIS WITH RESEARCH IN RESEARCH PROJECTS

SUMMARY

Keywords: Data Mining, Cluster Analysis, Research Projects

In this study, the application drafts created for the research projects carried out through the scientific research project management system in the Sakarya University information system and the applications of the applications made were examined and multiple analyzes were performed by clustering according to the project study areas.

Before the acceptance process of the applications of previous research projects took place or even after the evaluation process, there was no healthy data record of the projects that were rejected. In 2018, all phases of project drafts and applications for research projects are recorded with the creation of projects and application phases in virtual environment. During the application of the project, information about the basic and science fields are requested when registering the keywords by the project manager. The support provided by the institution regarding the working areas of the applications and the model that determines the future investment areas are shown.

With the cluster analysis and the study areas of the research projects, the K-Means, X-Means and EM (Expectation Maximization) algorithms are applied for cluster analysis from data mining techniques and the algorithms are evaluated.

BÖLÜM 1. GİRİŞ

Bilgi insan beyninin idrak edebileceği gerçek, varlığı deneylerle kanıtlanmış ya da temel kuralların tümüdür. Bilgi yaşamın her safhasında kullanılır ve bilgiye ihtiyaç duyulur. Teknolojinin gelişmesi ile birlikte bilgiyi elde etme yöntemleri de gelişmiş ve bilgiyi elde etmek için yeni teknikler kullanılmıştır. Veri madenciliği bu tekniklerden biridir.

Veri madenciliği tekniğinin uygulanma alanları bir hayli geniştir. Bu alanlardan veri görselliği, veri tabanı sistemleri, istatistik, vb. gibi akademik alanlar olarak gösterilebilir (Savaş, Topaloğlu, Yılmaz, 2012).

Veri madenciliği ile büyük boyutlu verilerden faydalı bilgiye ulaşım sağlanabilir ve gelecek ile ilgili tahminlerde bulunmamıza yol açabilir. Ortak özelliklere sahip verilerin gruplanması ile faydalı desenler elde edilebilir. Bu, günümüzü ve geleceği yorumlama da katkı sağlayacak veri madenciliği tekniklerinden biri de Kümeleme Analizi'dir.

Araştırma projeleri, sonuçlandığında bilime ve insana çıktıları ile katkı sağlayan projelerdir. Bu projelerin tamamlanması ile toplumun refah seviyesi teknolojik, ekonomik ve insani değerler olarak yükselmesi beklenmektedir (Yök, bap yönetmeliği).

Kamu kurumlarında yıllık gelirlerin belirli bir kısmı araştırma projeleri için ayrılmaktadır. Ayrılan bütçe için harcama alım kalemlerinde belirli oranda bir dağılım söz konusudur yani kurumun ayırdığı gelir tüm harcama türleri için eşit değildir. Harcama işlemlerindeki dağılım kamu çıkarları gözetilerek Hazine ve Maliye Bakanlığı tarafından gerçekleştirilmektedir. Harcamalara ait mali kalemlerde belirli

bir ödenek sınıflandırması ya da gruplandırması varken, araştırma projelerine ait çalışmaların anahtar kelimelerinde de temel alan ve bilim alanı gibi belirleyici faktörler vardır ama bunların aktif belirleyici rolleri olmamıştır.

Bu çalışmada, bilgiler Sakarya Üniversitesi Bilgi Sistemi (SABİS)'de yer alan Bilimsel Araştırma Proje Yönetim Sistemi üzerinden yapılan araştırma projeleri başvurularını kapsamaktadır. Araştırma projelerinin etkin çalışma alanları eğilimi ve projelerin temel alanları ve bilim alanları baz alınarak analiz edilmiştir.

Bu bölümde çalışmanın temel konusu, kaynağı açıklanırken, ikinci bölümde kaynak araştırması yapılmıştır.

Üçüncü bölümde, doğru bilgiye ulaşmak için yapılan veri madenciliği analiz yöntemlerinden biri olan kümeleme analizi kavramı anlatılmış ve uygulama için gerekli safhalar belirtilmiştir. Araştırma projelerine ait çalışma alanları anahtar kelimeler üzerinden elde edilen temel alan ve bilim alanları verileri üzerinde yapılan çalışmalar detaylandırılarak gösterilmiştir.

İgili sistem üzerinden elde edilen veriler kullanılarak veri madenciliği yöntemlerinden biri olan kümeleme analizi gerçekleştirilmiştir. Elde edilen bilgiler üzerinden aktif kullanılabilecek, kuruma öngörü sağlayabilecek bulguları elde etmek ve yeni model önerisi sunmak amaçlanmıştır. Bu öneriler ile ilgili kurum çalışma faaliyet alanlarının geleceğini nasıl etkileyebileceği ve yapılan çalışma alanlarının etkin adil bir mali dağılım ile ülkeye katkısının ne türlü sağlayabileceğini gözlemlenebilecektir.

Dördüncü bölümde ise çalışmanın uygulama sonuçları ve varsayımları detaylı bir şekilde anlatılmıştır. Elde edilen bilgiler üzerinden çeşitli çıkarımlar yapılmıştır.

Son bölümde ise analizden çıkarılan bulgular ve gelecek projelerin yapısı ve çalışma alanları ile ilgili önerilerde bulunulmuştur.

BÖLÜM 2. KAYNAK ARAŞTIRMASI

Yaşamın olduğu her yerde ve insanın var olduğu her safhada bilgiye ulaşmak ve bilgiyi kullanmak etkin önem arz eder. Çağımızda verilerin hızlı bir şekilde artması ve bu verilerin boyutlarının yüksek oranda hacim kaplaması ile bilgiye sağlıklı bir şekilde erişim için veri madenciliğine olan talep artmaktadır. Veri madenciliği teknikleri ile elde edilen bilgiler insana ve insan tarafından oluşturulan mekanizmalara ya da süreçlere değer katmaktadır.

Verilerin belirli gruplara veya sınıflara ayrılması ile veri madenciliği tekniklerinden kümeleme yöntemi uygulanarak sağlıklı analizler ve bulgular çıkarılabilir.

2019 yılında Kütahya Dumlupınar Üniversitesi'nde, bilimsel araştırma projesi hazırlamada fen bilimleri öğretmenlerinin bilimsel danışmanlık süreçleri incelenmiştir. Yapılan veri analizi sonuçlarında, proje hazırlama sürecine ilişkin öğretmenlerin bilimsel danışmanlık sürecine ilişkin yapılan uygulamaların öğretmenlerin mesleki gelişimine olumlu katkı sağladığı ve proje çalışmalarında rehbere ihtiyaç duyulduğu sonucuna ulaşılmıştır (Çetintaş, 2019).

Gazi Üniversitesi'nde yapılan akademik araştırma projelerinin değerlendirilmesine yönelik çalışmada, belirli araştırma projeleri üzerinden analitik hiyerarşik süreci (AHP) kurularak, projelerin çok kriterli karar verme (TOPSIS) algoritması ile kriter değerlendirilmesi yapılmıştır. Belirli ölçüler çerçevesinde en iyi araştırma projeleri hızlı bir şekilde belirlenmektedir (Arıbaş, 2016).

Çanakkale Onsekiz Mart Üniversitesi tarafından 2002 - 2009 yılları arasında desteklenen bilimsel araştırma projelerinin etkinliği ve mali yapısının incelenmesi ile üniversitenin finansman yapısının değişmesi ve toplam gelirleri içerisinde döner

sermaye gelirleri artarken, kamu bütçesinden aldıkları payların azaldığı sonucuna ulaşılmıştır (Boz, 2011).

Marmara Üniversitesi'nde yapılan araştırma geliştirme faaliyetleri üzerine ele alınan üniversite sanayi işbirliği bilimsel araştırma projeleri değerlendirilerek, sanayi ve üniversiteler arasındaki bağın zayıflığı bulgularına ulaşılmış ve aralarındaki bağı kuvvetlendirme üzerine öneriler sunulmuş ve bu öneriler ile Türk ekonomisinin diğer ülkelerle rekabet edebilme yeteneğini geliştirebileceği vurgulanmıştır (Arıcan, 2010).

Gazi Üniversitesi Bilimsel Araştırma Projeleri Müdürlüğü için 2004 yılında bir karar destek sistemi oluşturulmuştur. Yapılan bu çalışmanın amacı, projelerin istatistiksel ve birçok kriterin göz önüne alındığı bir sistem ile müdürlük tarafından alınan proje kararlarının en objektif ve verimli projelerin desteklenmesini sağlamaktır (Akaner, 2004).

Düzce Üniversitesi'nde yapılan bir çalışmada Weka programı ile lenf kanseri verileri kullanılarak kümeleme analizi yapılmıştır. Kümeleme algoritmaları olan; K-means, Hiyerarşik ve Em algoritmalarının performansları değerlendirilerek K-means algoritmasının diğer algoritmalara göre daha yüksek performansa sahip olduğu sonucuna ulaşılmıştır (Aksakallı, 2019).

2019 yılında Selçuk Üniversitesi'nde yapılan kümeleme analizi yarasa algoritması ile çeşitli tıbbi veri setleri üzerinden gerçekleştirilmiş ve 30 uygulamadan bazılarında veri örneklerinin hemen hemen %78'nin yarasa algoritması ile doğru kümelenebildiği görülmüştür (Fadhil, 2019).

Afyon Kocatepe Üniversitesi'nde, kütüphane verileri üzerinde veri madenciliği yöntemlerinden biri olan kümeleme analizi gerçekleştirilmiştir. Weka programı ile Farthest First, Filtered Clusterer, Make Densit Based Clusterer ve Simple K-Means kümeleme analizi yöntemleri uygulanarak, öğrencilerin kitap seçimleri üzerindeki davranışları, bölümlere göre kitap tercihleri grafiklerle gösterilmiştir (Gürel, 2019).

Yıldız Teknik Üniversitesi'nde yapılan bir çalışmada, hiyerarşik olmayan kümeleme yöntemlerinden biri olan K-Ortalamalar algoritması kullanılarak, Migros Ticaret A.Ş.'ne ait müşteri verileri üzerinden satın alma davranışları analiz edilmiştir. Çalışma sonucunda müşteri tercihleri dikkate alınarak, firma için talep oluştururken bir yandan da doğru talebi doğru zamanda müşteriye sunma sağlanmıştır (Üstünel, 2018).

Ankara Üniversitesi'nde, Mezotelyoma hastalığı teşhisi üzerinde yapılan çalışmada gerçek hasta verileri ile kümeleme yöntemlerinden Izgara Bölümleme, Alt Kümeleme ve Bulanık-C Ortalamalar kullanılmış ve en iyi sonuçların Bulanık-C Ortalamalar (Fuzzy-C Means) ile alındığı görülerek maliyet, zaman ve verilen kararların hata oranını azaltmak amaçlı karar destek sistemi oluşturulmuştur (Kaya, 2018).

İstanbul Aydın Üniversitesi'nde, dünya bankasının web sitesinden 2015 yılına ait 214 ülkenin verilerinden yararlanılarak, kümeleme analizi yöntemlerinden olan K-Means ve Self Organizing Map algoritmaları uygulanarak, ülkelerin gelişmişlik ölçüsüne göre kümelenmesi ile Türkiye'nin oluşan kümelerin içerisindeki yeri incelenmiştir (Akkuş, 2017).

Erciyes Üniversitesi'nde mikrodizi verilerinin analizinde, genlerin benzerlikleri ve farklılıklarının incelenmesi ve veriler arasında gruplandırma için veri madenciliği yöntemlerinden biri olan kümeleme analizi yapılmıştır. Kümeleme algoritmalarından K-ortalamalar (K-Means), kendi kendini düzenleyen haritalar (Self-Organizing Maps - SOM) ve veriye göre Adaptif DBSCAN (Adaptive DBSCAN According to data - ADBSCAN-ATD) algoritmaları kullanılmış ve ADBSCAN-ATD algoritmasının diğer algoritmalara göre daha verimli olduğu sonucuna ulaşılmıştır (Ulaş, 2016).

Bahçeşehir Üniversitesi'nde bir telekomünikasyon firmasına ait veriler istenilen formata getirmek için .net kullanılarak temizlenmiş, SPSS uygulaması ile güvenilirlik analizi yapılıp Weka uygulaması ile birliktelik kuralları çıkarılarak kümeleme analizi yapılmıştır. 25869 üye verisi üzerinden oluşan içerik türleri gruplarından en fazla komedi ve drama içeriğinin tercih edildiği gibi kurallar çıkarımı yapılmıştır. Verilerin belirli zaman aralığında alınması ile elde edilen birliktelik kurallarının, verilerin

süreklilik arz ettiği ortamda daha tutarlı ve kesin sonuçlar verebileceği belirtilmiştir (Tekingöz, 2016).

Atatürk Üniversitesi'nde yapılan bir çalışmada bir üniversiteye ait 63927 kütüphane verisi kullanılarak XLSTAT, R, Weka ve Rapid Miner (YALE) uygulamaları üzerinden kümeleme yöntemlerinden hiyerarşik K-Medoids, DBSCAN ve hiyerarşik olmayan K-Ortalamalar algoritmaları ile analiz yapılmıştır. Çalışma da Rapid Miner, Weka'ya göre daha ayrıntılı sonuçlar vermiştir (Gökalp, 2014).

Marmara Üniversitesi'nde Matlab uygulaması ile web haber sayfası dökümanları veri setleri üzerinden bölünmeli kümeleme yöntemlerinden K-Means, K-Medoids ve Fuzzy C-Means algoritmaları, beş farklı veri seti üzerinden kümeleme performans karşılaştırmaları yapılmıştır. Sonuç olarak K-Medoids algoritması ile en iyi kümeleme işlemi gerçekleştirilirken en hızlı sonuç K-Means algoritması ile elde edilmiştir (Işık, 2006).

Milli Savunma Üniversitesi'nde, Meteoroloji Genel Müdürlüğü'nden sağlanan veriler ile Türkiye'deki illere üzerinden 1545 gözlem istasyonuna ait 9 ayrı meteorolojik veri kullanılarak Rapid Miner, Weka ve SPSS uygulamaları ile kümeleme çalışması yapılmıştır. İllerin meteorolojik olarak sınıflandırılması ile çeşitli kurum ve kuruluşların yatırım planlarına yön vermede yardımcı olması amaçlanmıştır (Kılınç, 2019).

Yıldız Teknik Üniversitesinde yapılan bir çalışmada Namık Kemal Üniversitesi Çorlu Mühendislik Fakültesi 2011 yılı sonu itibari ile kayıtlı lisans öğrencilerinin verileri üzerinden kümeleme analizi teknikleri Enterprise Miner, Weka ve Matlab uygulamaları ile gerçekleştirilmiş ve Bilgisayar Mühendisliği öğrencilerileri ile 271 anket çalışması yapılarak başarılarındaki aile faktörü etkisi araştırılmıştır (Saygılı, 2013).

Veri madenciliğinde yapılan kümeleme çalışmalarının ortak paydaları farklı metotlar kullanarak verilerin daha hızlı, etkin ve güvenilir bir şekilde analiz edilmesidir.

BÖLÜM 3. MATERYAL VE YÖNTEM

3.1. Kümeleme Analizi

3.1.1. Kümeleme analizi nedir?

Kümeleme Analizi, veri veya veri setlerinin benzer özelliklerine göre sınıflara veya gruplara ayırmak için kullanılan istatistiksel analiz tekniğidir. Kümeleme analizi ile oluşturulan kümeler aynı küme içerisinde yer alan veriler diğer kümelere göre daha çok benzer özellik göstermektedir (Hair., 1998:473).

3.1.2. Kümeleme analizinin amacı

Değişik tür de verilerin bir araya getirilerek veri madenciliği tekniklerinden birisi olan kümeleme analizi ile sınıf sayısı bilinmeyen ve sınıflandırılmamış verilerin benzer özelliklerine göre sınıflandırılması amacıyla kullanılmaktadır. Kümeleme analizi verilerin türlerine veya çeşitli özelliklerine göre birbirlerine benzer ortak yanları ortaya konularak farklı kümelerde bir araya getirilmesini sağlayan bir tekniktir. Kümeleme analizi ortak özellikleri olan verilerin ya da nesnelerin aynı sınıflarda bir araya getirilmesini amaçlaması bakımından diskriminant analizi ile, birbirleri ile ortak özellikleri olan değişkenlerin aynı sınıflarda toplanmasını amaçlaması sebebiyle de faktör analizi ile benzeşim göstermekte olup, veri indirgeme hususiyeti vardır (Çakmak, 1999:s.188).

3.1.3. Kümeleme analizinin kullanım alanları

Araştırma problemlerinde geniş yer kaplayan kümeleme analizi özellikle sağlık alanında hastalıklarla ve hastalara uygulanan tedavilerin, sınıflandırılarak

kümelenmesi ile çok yaygındır. Hastalık türlerinin ayırt edici özellikleri ile onlara uygulanarak (Bircan, H., ve Çam, S. 2016 ; Talan, M. İ. 2016 ; Azak, E. N., ve Şen, H. 2019) elde edilen bulguların gruplandırılması tedavi süreçlerinin hızlanmasına sebep olmuştur.

Sosyal bilimler ve fen bilimlerinde de kümeleme analizi teknikleri uygulanmaktadır. Sosyal bilimlerde ülkeleri ya da şehirleri sosyal ve ekonomik açıdan incelerken fen bilimlerinde son yıllarda veri madenciliği ile büyük verilerden özet veriler elde etmek için kümeleme analizi kullanılır (Gürbüz ve Karabulut, 2009 ; Koldere Akın, 2008).

3.1.4. Kümeleme analizi benzerlik ve uzaklık ölçüleri

Veri setinde yer alan verilerin veri tipi göz önüne alınarak verilerin birbirlerine olan benzerlikleri ya da verilerin birbirlerine olan uzaklıkları ile kümeleme işlemi gerçekleştirilir. Elde edilen değişkenlerin yapısına göre hangi benzerlik ölçüsünün ya da uzaklık ölçüsünün kullanılacağına karar verilir. Eğer değişkenler farklı ölçüm birimleri olduğu ve standart sapma ile ortalamanın birbirlerine göre çok farklılık göstermesi benzerlik ve uzaklık ölçülerinin yapılmadan önce verilerin ölçülebilir hale getirilmesi gerekir.

3.1.5. Aralık ölçekli ve oransal ölçekli değişkenler için uzaklık ölçüleri

3.1.5.1. Öklidyen uzaklık ölçüsü

En sık kullanılan uzaklık ölçülerinden biri Öklidyen Uzaklık ölçüsü ya da onun karesidir. Aşağıdaki formül ile öklidyen uzaklığı hesaplanırken,

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2} \quad (3.1)$$

öklidyen uzaklığının karesi de,

$$d(i,j) = (x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2 \quad (3.2)$$

formülü ile hesaplanır.

Öklid uzaklığı hesaplanırken ölçülü veriler üzerinde işleme gerçekleşmemiş ham veriler üzerinde hesaplama yapılır. Bu nedenle yeni verilerin eklenmesi kümeleme analizini etkilemez (Demiralay, 2005).

3.1.5.2. Pearson uzaklık ölçüsü

İki birim arasındaki uzaklık ölçüsü ve

$$(i,j) = \sqrt{\frac{(x_{i1}-x_{j1})^2}{S_1^2} + \frac{(x_{i2}-x_{j2})^2}{S_2^2} + \cdots + \frac{(x_{ip}-x_{jp})^2}{S_p^2}} \quad (3.3)$$

formülü ile Pearson uzaklık ölçüsü hesaplanır. Bu formülde kullanılan S'ler varyans tır, ilgili özelliğe ait muhtemel standart sapmadır.

3.1.5.3. Manhattan uzaklık ölçüsü

Manhattan bir diğer tabir ile şehir bloğu uzaklık ölçüsü, gruplar arasındaki uzaklık ölçüsünün mutlak değerlerinin toplamı ile bulunur. Formülü;

$$d(i,j) = \left(|x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|\right) \quad (3.4)$$

ile hesaplanır.

Ölçüm yapılan değişkenlerin türleri farklı olması ile standartlaştırılmış Öklid uzaklığının karesi ile karşılaştırıldığında Manhattan uzaklık ölçüsünün yorumlanabilir sonuçlar vermediği görülebilmektedir (Atbaş, 2008).

3.1.5.4. Minkowski uzaklık ölçüsü

Minkowski uzaklık ölçüsü diğer formüllerin genelleştirilmiş halidir.

$$d(i,j) = \left[\left|x_{i1} - x_{j1}\right|^m + \left|x_{i2} - x_{j2}\right|^m + \cdots + \left|x_{ip} - x_{jp}\right|^m\right]^{1/m} \quad (3.5)$$

u uzaklık ölçüsündeki m değeri değişen farklara verilen ağırlığı değiştirir. m=1 olarak alındığında formül Manhattan uzaklık ölçüsünün formülüne, m = 2 olarak alındığında formül Öklid uzaklık ölçüsü formülüne dönüşür (Anderberg 1973).

3.1.5.5. Mahalanobis uzaklık ölçüsü

Mahalanobis uzaklık ölçüsü iki birim arasındaki kovaryans veya korelasyonu ön plana alır ve birimler arasındaki uzaklığı doğrudan birleştirme yaparak hesaplarken

$$d(i,j) = D^2 = \left(x_i - x_j\right)' S^{-1}(x_i - x_j) \quad (3.6)$$

formülünü kullanır.

Mahalanobis uzaklık ölçüsü aykırı birimleri hesaplaması ile avantaj oluşturur ve bu yönüyle uzaklık ölçüleri arasında en avantajlısı Mahalanobis uzaklık ölçüsüdür denebilir (Sharma 1996).

3.1.6. Kümeleme analizi yöntemleri

Kümeleme, soyut veya somut varlıkları ortak ya da benzeri varlık sınıfları içerisinde gruplama sürecidir (Kaufman ve Rousseeuw, 1987). Eldeki veriler üzerinde hangi uzaklık ölçüsü kullanılacağına karar verildikten sonra ilgili kümeleme yöntemine karar verme aşamasına geçilir.

Benzerliklerine göre kümelere dahil edilerek uygulanabilecek çeşitli yöntemler vardır (Dinler, 2014). Tüm yöntemler içinde en önemlisi, kümelerin içindeki benzerliklerin en üst seviyede olurken kümeler arası farkların da en üst seviyede olmasıdır. Kümeleme algoritmaları, en çok kullanılan hiyerarşik ve hiyerarşik olmayan kümeleme yöntemleri olarak iki temel yapıya ayrılarak sınıflandırılmıştır (Blashfield ve Aldenferder 1978 ; Özdamar 2004).

3.1.6.1. Hiyerarşik kümeleme (aşamalı) yöntemleri

Veri analizinde küçük örneklemler için hiyerarşik kümeleme analizi uygun görülmektedir (Çokluk vd., 2016: 142). Aşama sıralı bir yöntem olarak bilinen hiyerarşik kümeleme yöntemi, gruplayıcı ve bölücü olarak kendi içerisinde ikiye ayrılır.

- Gruplayıcı Hiyerarşik Yöntem

Veri setinde yer alan her bir birimin ilk olarak ayrı bir küme olduğu düşünülerek kümeleme işlemine başlanır ve kümeler arası benzerlik seviyelerine göre en yakın iki küme yeni bir küme de toplanır. Böylelikle her işlemde küme sayısı azalır. Hiyerarşik kümeleme tekniği ile birleşen gruplar daha sonraki işlem adımlarında ayrılamazlar (Fırat, 1997).

- Tek Bağlantı Metodu

Diğer adı ile en yakın komşu bağlantı kümeleme yöntemi olarak geçen tek bağlantı metodu tekniğinde birbirine en yakın kümeler aralarındaki mesafeler uzaklık matrisi ile belirlenerek birleştirilir (Albayrak, 2019). Tek bir küme kalana kadar aynı işlem tekrar eder.

- Tam Bağlantı Metodu

Tam bağlantı metodunda tek bağlantı metodunda olduğu gibi işlemler gerçekleştirilir tek farkı vardır, en uzak kümeler birleştirilerek işlemler yapılır (Green 1989). Bu metoda en uzak komşu bağlantı kümeleme yöntemi de denir (Günay Atbaş, 2008).

- Ortalama Bağlantı Metodu

Bu bağlantı metodunda tek bağlantı ve tam bağlantı metotlarında olduğu gibi işleme başlanır yalnız kümeleme işlemi küme içerisinde yer alan birimlerin birbirlerine

ortalama uzaklığı baz alınarak gerçekleştirilir. Sonuçları ise tek ve tam bağlantı yöntemlerinin bulgularına ortalama değer vermesi ile alternatif bir yöntem olarak tercih edilir (Hubert, 1974).

- Merkezi Metodu

Küme merkezinin değeri kümeyi oluşturan birimlerin ortalaması alınarak hesaplanır, eğer küme tek bir birimden oluşuyorsa merkez o birim olarak kabul edilir.

- Varyans (Ward's) Metodu

Bu metod bir kümenin grup içi bağlantılarının ortalama uzaklık değerini baz alarak varyansın minimum olmasını amaçlar (Dahl ve Naes, 2004; He, 1996). Kareler toplamından faydalanılan bu yöntem ile en az bilgi kaybı yaşanır (Akat, 2007).

- Bölücü Hiyerarşik Yöntemler

Gruplayıcı hiyerarşik yöntemlerin süreçlerine göre aksi bir yöntem uygulanır. Tüm birimlerin dahil edildiği bir küme üzerinden birbirine uzak olan birimlerin ayrılması ile yeni kümeler oluşturulur. Her birimin tek bir kümeyi oluşturması safhasına dek işlemlere devam edilir (Everitt ve ark., 2011).

3.1.6.2. Hiyerarşik olmayan (aşamalı olmayan) yöntemler

Analiz öncesi küme sayısında ön bilgi mevcut ise ya da araştırmacı tarafında küme sayısı daha önceden belirlenmiş ise hiyerarşik olmayan kümeleme yöntemi tercih edilir (Tatlıdil, 2002: 338). Hiyerarşik olmayan birçok kümeleme yönteminden bahsedilebilir ancak en sık kullanılan K-Ortalamalar ve En Çok Olabilirlik yöntemleridir.

- K-Ortalama Metodu

K-Ortalamalar en çok tercih edilen bir kümeleme yöntemi olup aşamalı olmayan bir yapıya sahiptir (Burn ve ark., 1997 ; Lin ve Chen, 2006). Bu yöntem her bir birim ile ona en yakın merkez arasında kalan Öklit mesafesinin değerinin tamamını en düşük seviyeye getirerek mevcut veri setini ayrı kümelere ayırmaktadır (Fırat ve ark., 2012).

- En Çok Olabilirlik Metodu

Diskriminant analizinde de tercih edilen en çok olabilirlik yönteminde her bir birim en yüksek olabilirlik değerini sonuç olarak verecek şekilde önceden belirlenen kümelere atanır. En çok olabilirlik yöntemi yaygın olarak kullanılmasa da kurumsal dayanağı güçlüdür.

3.2. Araştırma Projeleri

Proje, bir araştırmanın ilerleme adımlarından oluşur (Avcı ve ark., 2016). Bilimsel Araştırma Projeleri ülkenin teknolojik, ekonomik, sosyal ve kültürel kalkınmasına katkı sağlayan bilimsel projelerdir. Bu projeler sonuçları itibariyle hem bilimsel ilerlemeye hem de Üniversitelerin kurumsal yapısına etki etmektedir. Üniversitelerde ve/veya özel/kamu sektör işbirliği ile de gerçekleştirilebilen bu projeler ile bilim insanı yetiştirmek, araştırma altyapısı kurmak ve geliştirmek amaçlanmaktadır.

BÖLÜM 4. ARAŞTIRMA BULGULARI

Makine öğreniminde özellikle karar destek sistemlerinde yaygın olarak kümeleme analizi ve algoritmaları kullanılır. Araştırma projelerine ait anahtar kelimeler ile ilgili kurumun projeler üzerinde çalışma alanları tespiti ve yapılan analiz sonuçları ile gelecek çalışma alanlarının ya da zayıf kalan çalışma alanlarının tespiti yapılabilecektir. Modelin oluşturması için çalışma esnasında Şekil 4.1.'de yer alan işlemler gerçekleştirilmiştir.

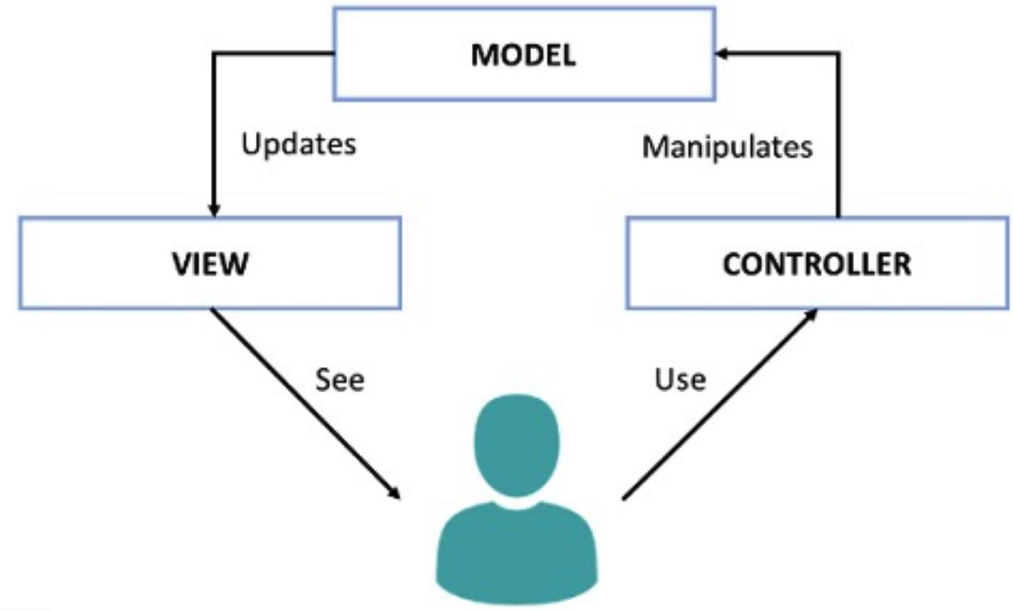
Şekil 4.1. Araştırma projelerinde çalışma alanları tespiti için iş aşamaları

4.1. Araştırma Projeleri

Tamamlandığında bilime, topluma, ülkeye ve ekonomiye katkı sağlayan araştırma projeleri, çalışma alanları ile yapılan araştırmaların seyrinin hangi doğrultuda ilerlediği bu tez çalışmasını kapsamaktadır. Sakarya Üniversitesi, Bilimsel Araştırma Proje Yönetim Sistemi oluşturularak, araştırma projelerinin başvuruları web ortamında alınarak gerekli veriler temin edilmiştir.

4.1.1. Bilimsel araştırma proje yönetim sistemi

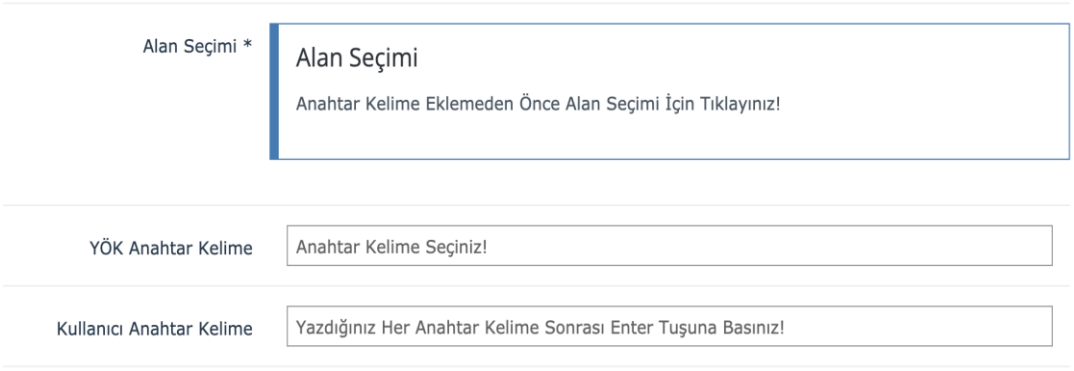
Bilimsel araştırma proje yönetim sistemi, asp.net MVC (Model-View-Controller) Framework ile yazılımı gerçekleştirilen 3 katmanlı bir sistem model yapısına sahiptir.



Şekil 4.2. MVC model yapısı

4.1.2. Araştırma projeleri anahtar kelimeleri

Bilimsel araştırma proje yönetim sistemi üzerinden yapılan başvurulardan her proje başvurusu için kullanıcılardan projeye ait çalışma alanları doğrultusunda oluşturacakları anahtar kelimeler istenmektedir.



Şekil 4.3. Web ortamında proje başvurusunda proje yürütücülerinden istenen çalışma alan bilgileri ve anahtar kelimeler

Araştırma projelerine ait anahtar kelimeler, alan seçimi doğrultusunda oluşturulmaktadır. Alan seçimi; temel alan ve onun altında yer alan bilim alanının seçimi ile gerçekleştirilir. YÖK (Yüksek Öğretim Kurumu)'da tanımlı anahtar kelimeler kullanıcının seçimine sunulur. Sistem üzerinde tanımlı olmayan anahtar kelimeler için kullanıcıya projesine ait anahtar kelimeleri oluşturması için imkan sağlanır.

Tablo 4.1. Anahtar kelime temel alanlar

Sıra	Temel Alan Adı
1	Eğitim Bilimleri ve Öğretmen Yetiştirme
2	Fen Bilimleri ve Matematik
3	Filoloji
4	Güzel Sanatlar
5	Hukuk
6	İlahiyat
7	Mimarlık, Planlama ve Tasarım
8	Mühendislik
9	Sosyal, Beşeri ve İdari Bilimler
10	Ziraat, Orman ve Su Ürünleri
11	Spor Bilimleri
12	Sağlık Bilimleri

12 farklı temel alan mevcuttur, bu temel çalışma alanlarının altında 300 farklı bilim alanı vardır. YÖK bilim alanlarından ise 1606 anahtar kelime kullanıcıya sunulmuştur.

Tablo 4.2. Anahtar kelime bilim alanları

Sıra	Temel Alan	Bilim Alan Adı
1	1	Bilgisayar Öğretim Teknolojileri Eğitimi
2	1	Din Kültürü ve Ahlak Bilgisi Eğitimi
3	1	Eğitim Bilimleri
4	1	Güzel Sanatlar Eğitimi
5	1	Ortaöğretim Fen ve Matematik Alanlar Eğitimi
6	1	Ortaöğretim Sosyal Alanlar Eğitimi
7	1	Özel Eğitim
8	1	Rehberlik ve Psikolojik Danışmanlık
9	1	Temel Eğitim
...	...	...
...	...	...
...	...	...
298	12	Veterinerlik Virolojisi
299	12	Yoğun Bakım
300	12	Zootekni, Genetik ve Biyoistatistik

2018 yılında kullanılmaya başlayan ve günümüzde aktif kullanılan bilimsel araştırma proje yönetim sisteminde projeler için proje yürütücüleri tarafından YÖK tabanlı 1680 anahtar kelime oluşturulmuştur. Bu anahtar kelimelerden bazıları proje yürütücüleri tarafından pasif hale getirilmiştir. Proje yürütücülerinin projelerine ait ise YÖK anahtar kelimelerinden bağımsız 2950 anahtar kelime oluşturulmuş ve bunların içinden de bir kısmı yürütücü tarafından pasifleştirilmiştir.

Tablo 4.3. Projelere ait yök anahtar kelimeler

Sıra	Proje Nu.	Anahtar Kelime	Durum
1	1	482	True
2	1	482	False
3	2	209	True
4	4	209	True
5	5	1440	True
6	5	1442	True
7	6	1440	True
8	1	482	False
9	7	137	True
...	...	...	...
...	...	...	...
...	...	...	...
1378	1026	1393	False
1379	1027	714	True
1380	1027	718	True

Oluşturulan projeler üzerinden, proje yürütücüleri tarafından sisteme eklenen anahtar kelime sayısı 2950 adettir. Bunlardan kimileri aynı proje yürütücüleri tarafından

kullanımı pasifleştirilmiştir. Pasif olan anahtar kelime için ilgili tablo üzerinden durum bilgisi ‘False’ olarak gösterilmiştir.

Tablo 4.4. Projelere ait kullanıcı anahtar kelimeler

Sıra	Proje Nu.	Temel Alan	Bilim Alanı	Anahtar Kelime	Durum
1	1	8	62	Statik Var	False
2	1	8	62	Statik Var	False
3	4	4	32	Seramik	True
4	5	9	137	Kırım Savaşı	True
5	5	9	137	Osmanlı Devleti	True
6	5	9	137	Telgraf	True
7	5	9	137	Telgraf Hatları	True
8	6	9	137	Osmanlı Devleti’nde Ticaret	True
9	6	9	137	Marka Kültürü	True
10	6	9	137	Alamet-i Farika	True
...	...			...	...
...	...			...	...
...	...			...	...
2946	1031	0	0	Talazoparib	True
2947	1031	0	0	Toll Benzeri Reseptörler	True
2948	1032	12	188	Hemşire	True
2949	1032	12	188	Çocuk Sağlığı ve Hastalıkları	True
2950	1032	12	188	Yenidoğan	True

Proje yürütücüleri tarafından oluşturulan bazı anahtar kelimelerin alan bilgileri; temel alan ve bilim alanı seçimi yapılmadan kaydı oluşturulduğu gözlemlenmiştir.

4.2. Modelleme

Veri tabanında tamamlanmamış, veri kirliliğine sebep olan bir çok değer olabilir. Bunlara kullanılmayan/aktif olmayan anahtar kelimeleri dahil edebiliriz, ya da kullanıcı tarafından oluşturulan anahtar kelimelerde alan seçimi yapılmaması güvensiz sonuç almamıza sebep olabilir. Örneğin Tablo 4.3.’te temel alan ve bilim alanı seçilmemiş kullanıcı tarafından kaydı oluşturulmuş anahtar kelimeler yer almaktadır.

Verileri amacına uygun hale getirmek için veri madenciliğinde veri temizleme tekniği uygulanır. Bu aşamada eksik, tutarsız ya da sapan değeri azaltmak için ilgili veri tablolarında pasifleştirilen veriler ile alan seçimi yapılmayan anahtar kelimeler silinerek, araştırma projelerine ait proje yürütücüsü tarafından oluşturulan anahtar

kelimeler ve YÖK tabanlı seçimi yapılan anahtar kelimelere ait temel alan ve bilim alan bilgileri birleştirilerek bilgi sistemi oluşturuldu.

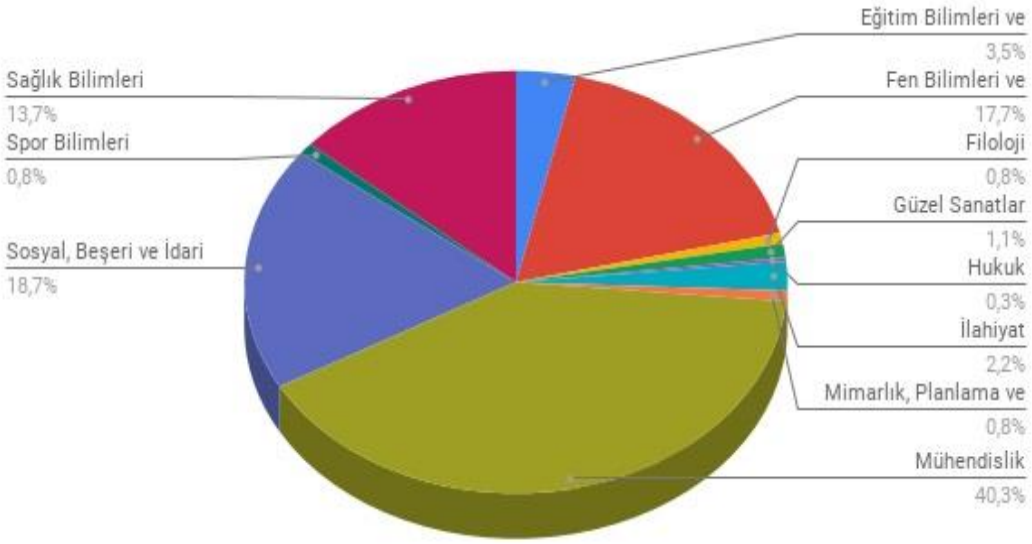
Veri temizleme işlemi ile YÖK tabanlı kaydı tutulan projelere ait anahtar kelime sayısı 1149'a düşerken, proje yürütücüleri tarafından oluşturulan anahtar kelime sayısı 2302'e düşmüştür. Toplamda oluşturulan tablo üzerinde 3551 anahtar kelimeye ait temel alan ve bilim alanı bilgileri yer almaktadır.

Tablo 4.5. Proje Anahtar Kelimelerine ait Bilgi Sistem Tablosu

Sıra	Proje Nu.	Temel Alan	Bilim Alanı
1	1	4	32
2	5	9	137
3	5	9	137
4	2	2	15
5	7	4	29
6	8	8	71
7	7	4	32
8	10	7	53
9	8	12	233
...		...	...
...		...	...
...		...	...
3549	1030	9	140
3550	1030	9	141
3551	1032	11	154

4.3. Araştırma Projelerinde Çalışma Alanları Kümelenmesi

Bu çalışma, veri madenciliği tekniklerinden biri olan kümeleme analizi ile araştırma projelerine ait anahtar kelimelerden oluşan bilgi sistemi Weka 3.9.4 uygulaması üzerinde K-Means, X-Means ve EM algoritmaları kullanılarak gerçekleştirilmiştir.



Şekil 4.4. Araştırma projelerinde temel alan bilgilerin dağılım grafiği

Temel alanların içerisinde 10. temel alan veri setinde yer almadığı gözlemlenmiştir. Ziraat, Orman ve Su Ürünleri ilgili fakültenin Sakarya Üniversitesinde yer almaması sebebine bağlanarak bir küme baz alınmamıştır.

4.3.1. Weka 3.9.4

Waikato Üniversitesi tarafından geliştirilen Weka uygulaması Java dili üzerinde geliştirilmiştir. Uygulamanın ismi Waikato Environment for Knowledge Analysis kelimelerinin baş harflerinden oluşur. Yaygın olarak bilgisayar bilimlerinden makine öğrenimi için kullanılmaktadır. Veri madenciliği adımlarından, veri ön işleme, ilkelleme, sınıflandırma ve gruplama gibi bir çok işlevin yanı sıra sonuçların görsel olarak gösterimine de olanak sunmaktadır (Şeker, 2020).

Şekil 4.5. Weka 3.9.4. uygulaması ana menü.

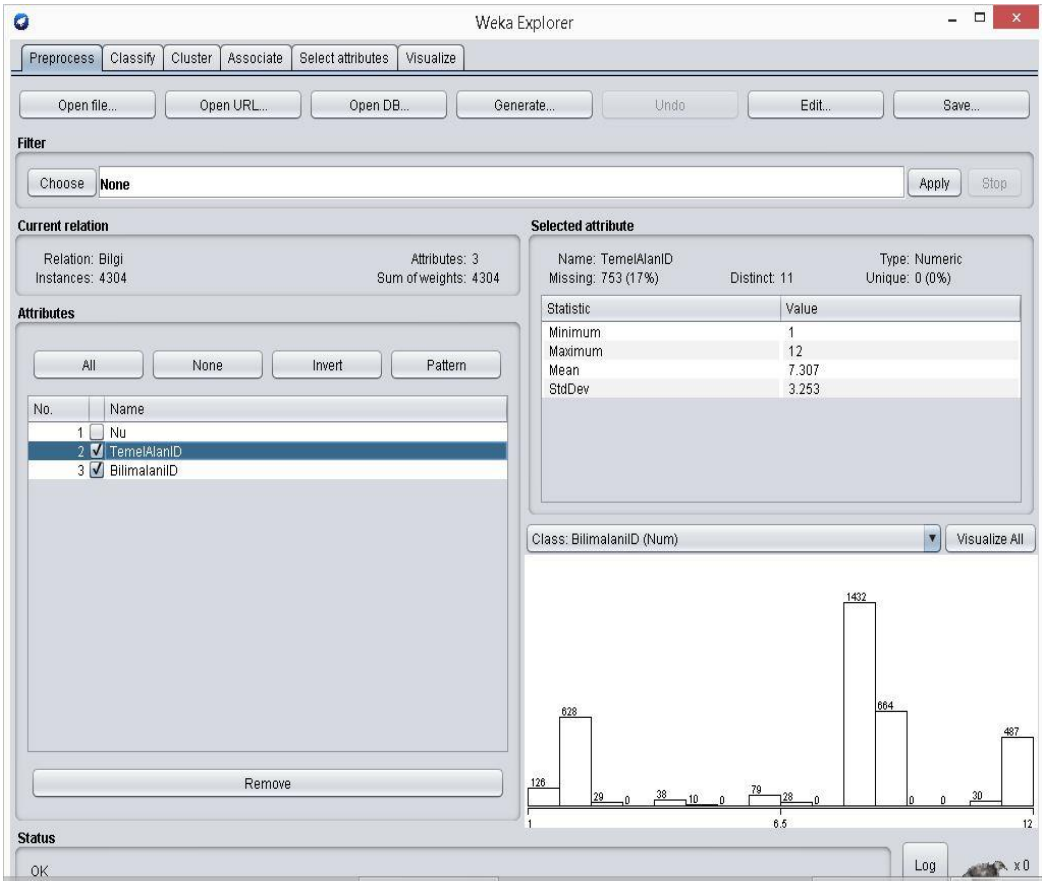
Veri ön işlemleri yapıldıktan sonra, araştırma projelerinde yer alan anahtar kelimelerden oluşan bilgi sistem tablosu Weka uygulamasına aktarabilmek için .csv dosyası oluşturulmuştur.

Viewer

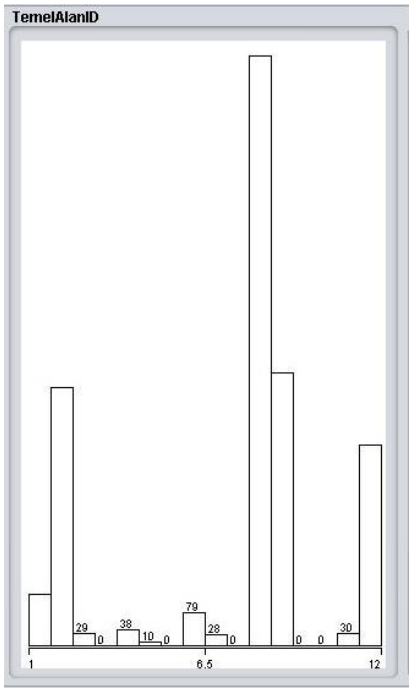
Relation: Bilgi Sistemi-BS

No.	1: ï»¿Nu	2: TemelAlanID	3: BilimalanID
	Numeric	Numeric	Numeric
1	1.0	4.0	32.0
2	2.0	9.0	137.0
3	3.0	9.0	137.0
4	4.0	9.0	137.0
5	5.0	9.0	137.0
6	6.0	9.0	137.0
7	7.0	9.0	137.0
8	8.0	9.0	137.0
9	9.0	9.0	137.0
10	10.0	9.0	137.0
11	11.0	2.0	15.0
12	12.0	2.0	15.0
13	13.0	2.0	15.0
14	14.0	2.0	15.0
15	15.0	2.0	15.0
16	16.0	4.0	29.0
17	17.0	4.0	29.0
18	18.0	4.0	29.0
19	19.0	4.0	29.0
20	20.0	4.0	29.0
21	21.0	4.0	32.0
22	22.0	8.0	71.0
23	23.0	8.0	71.0
24	24.0	8.0	60.0
25	25.0	8.0	60.0
26	26.0	9.0	140.0
27	27.0	9.0	140.0
28	28.0	9.0	140.0
29	29.0	9.0	140.0

Şekil 4.6. Weka üzerinde normalleştirilmiş veri seti

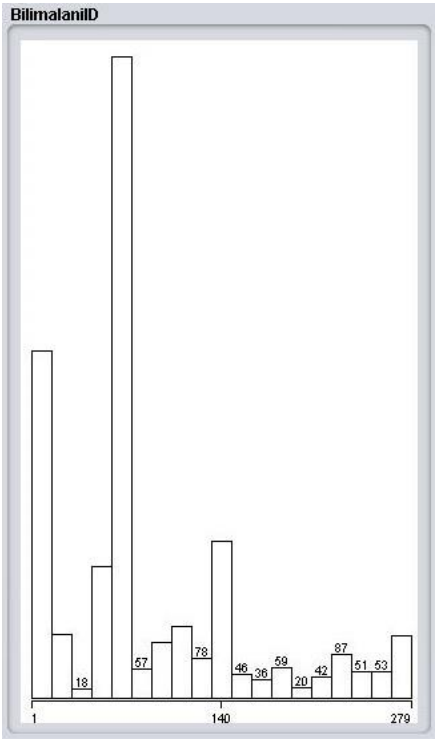


Şekil 4.7. Weka uygulaması üzerine aktarılan veri seti



Şekil 4.8. Anahtar kelimelere ait temel alan bilgisi kullanım grafiği

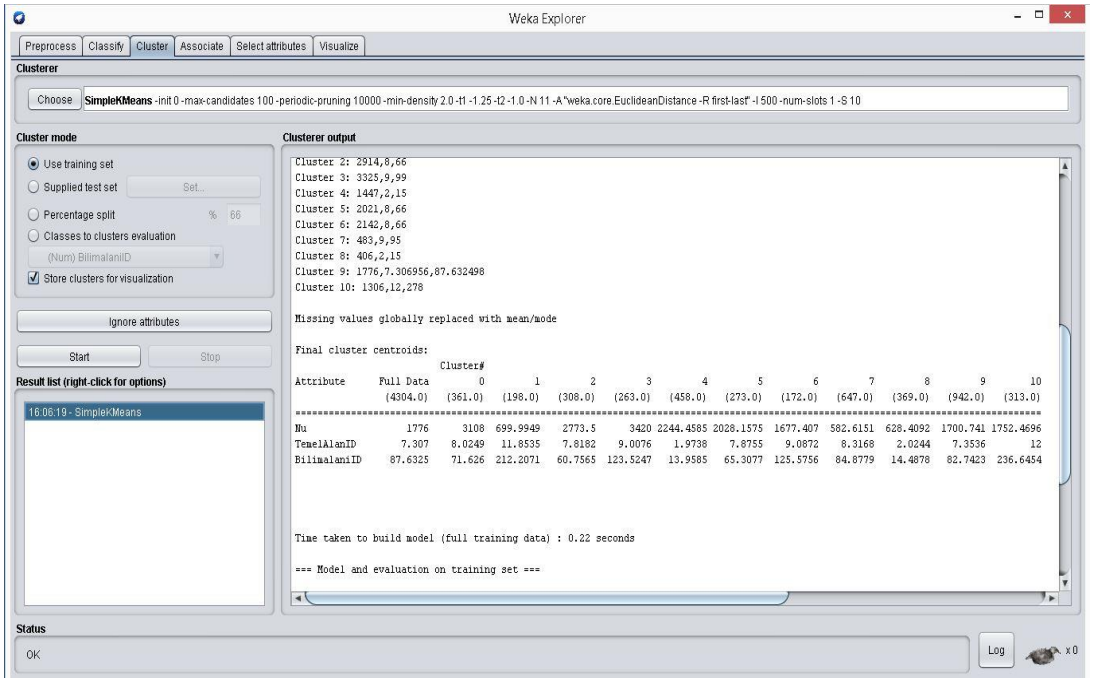
Uygulamaya alınan veri seti üzerinden Şekil 4.8.'de görüldüğü gibi anahtar kelimelere ait 12 farklı temel alan bilgisi ve bunların kullanım grafiği gösterilmektedir. Şekil 4.9.'de ise ilgili anahtar kelimelere ait bilim alanları kullanım grafiği yer almaktadır.



Şekil 4.9. Anahtar kelimelere ait bilim alanı kullanım grafiği

4.3.2. K-Means algoritması ile elde edilen sonuçlar

Sadece sayısal değerler ile çalışan K-Means kümeleme algoritması kullanımı için metinsel veriler yerine numerik bir veri seti oluşturulmuştur. Weka uygulaması üzerinde veri seti dosyası kullanımı için dosya türü .csv formatına getirilmiştir. Burada dikkat edilmesi gereken bir husus oluşturulan .csv dosyasında değişkenler arasında yer alan noktalı virgüllerin virgüle dönüştürülmesidir. Herhangi bir text editörü buna yardımcı olacaktır. Aksi durumda Weka uygulaması noktalı virgülle ayrılan değişkenleri algılayamacaktır. İlgili veri setine K-Means algoritması uygulanarak oluşturulacak küme sayısı 11 olarak belirlenmiştir, elde edilen dağılım Şekil 4.10.'da yer almaktadır.



Şekil 4.10. K-Means algoritması uygulanan veri setinin kümelenmesi

Küme dağılımlarının nominal değerlere sahip olduğu Şekil 4.10.'da görülmektedir ve bu sebeple normalize yapılmaya gerek duyulmamıştır. Kümeleme işlemi K-Means algoritmasında 0.22 sn süre içerisinde gerçekleştirilmiştir.

Kümeler üzerindeki dağılım en küçük küme yapısında %4 ile 172 birime sahip iken en büyük küme yapısı %22 oranı ile 942 birime Şekil 4.11.'da görüldüğü gibi sahiptir.

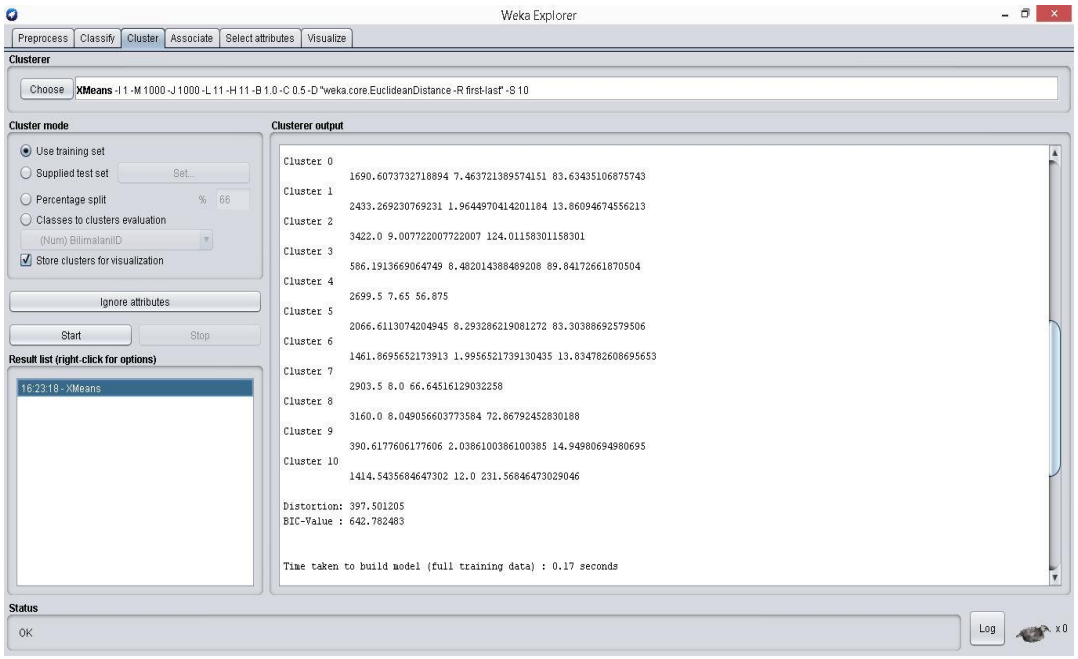
Clustered Instances

0	361	(8%)
1	198	(5%)
2	308	(7%)
3	263	(6%)
4	458	(11%)
5	273	(6%)
6	172	(4%)
7	647	(15%)
8	369	(9%)
9	942	(22%)
10	313	(7%)

Şekil 4.11. K-Means ile oluşan kümelere ait değerler

4.3.3. X-Means algoritması ile elde edilen sonuçlar

X-Means algoritması Weka uygulaması üzerinde hazır kütüphane paketi olarak gelmediği için, Weka uygulaması ana ekranın da yer alan tools menüsü üzerinden package manager sekmesinden yararlanılarak X-Means kütüphanesi uygulamaya yüklenmiştir. Bu algoritmanın kullanımında en büyük ve en küçük küme sayısı kullanıcı tarafından belirlenmektedir, nominal veri kabul etmemekle birlikte veri yapısı kendisine özeldir. İlgili veri seti uygulama üzerinden seçilerek X-Means kümeleme algoritması çalıştırılarak işlemler 0.17 saniye de gerçekleştirilmiştir.



Şekil 4.12. X-Means algoritması uygulanan veri setinin kümelenmesi

Kümeleme işlemi için minimum küme değeri K-Means algoritmasında olduğu gibi 11 olarak belirlenmiştir. Bunun sebebi hiyerarşik olmayan kümeleme algoritmalarında küme sayısı belirlenmesinin kullanıcıya açık olması ve anahtar kelimeler üzerinde kullanılan temel alan sayısının 11 olması ile ilgilidir.

Clustered Instances

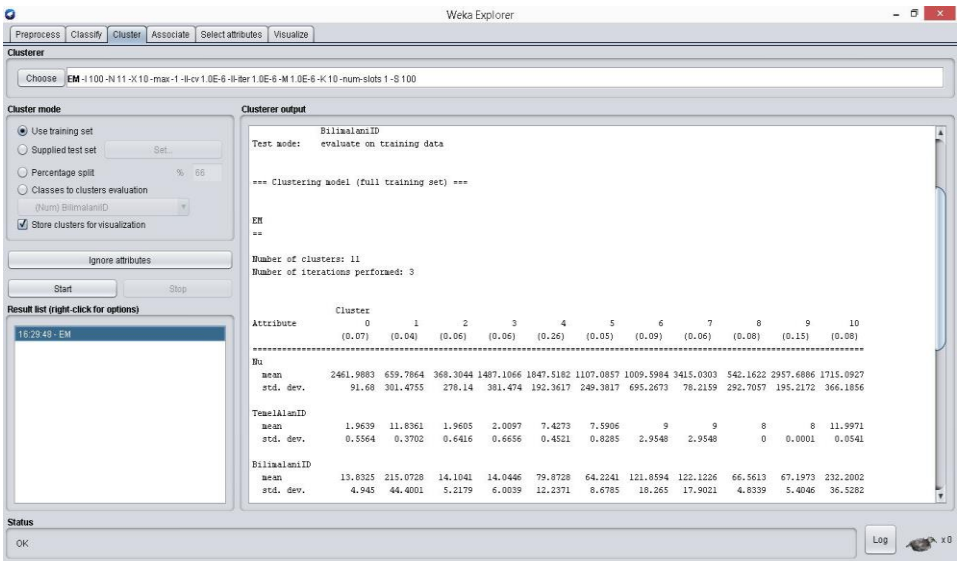
0	1085	(25%)
1	338	(8%)
2	259	(6%)
3	695	(16%)
4	160	(4%)
5	283	(7%)
6	230	(5%)
7	248	(6%)
8	265	(6%)
9	259	(6%)
10	482	(11%)

Şekil 4.13. X-Means ile oluşan kümelere ait değerler

X-Means algoritmasında oluşan kümeler üzerindeki dağılım en küçük küme değerine sahip olan %4 lik yapısı ile 160 birime sahip iken en büyük küme değerine sahip %25 lik oran ile 1085 birime sahip kümedir.

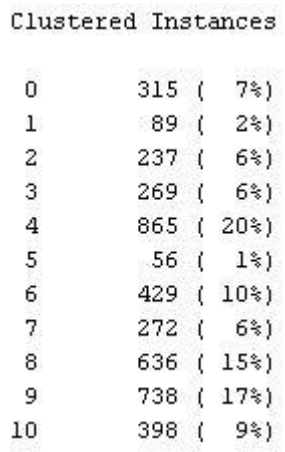
4.3.4. EM (Expectation Maximization) algoritması ile elde edilen sonuçlar

EM algoritması her tekrar ettiğinde olasılık fonksiyon sayısı da artar ve kayıp verilerinin kestirim değerini de hesaplar.



Şekil 4.14. EM algoritması uygulanan veri setinin kümelenmesi

Weka uygulaması üzerinde hazır kütüphane olarak yer alan EM algoritması küme değeri 11 olarak belirlenerek veri setine uygulanmıştır. Kümeleme işlemi 0.98 sn de gerçekleştirilmiştir.



Şekil 4.15. EM algoritması ile oluşan kümelere ait değerler

EM algoritması ile oluşan kümelerden en küçük yapıya sahip 56 birimi olan ve %1 lik oranı ile 5. küme iken en büyük küme yapısına sahip 4. küme olup %20 lik oran ile 865 birimi mevcuttur.

4.3.5. K-Means, X-Means ve EM algortimalarının karşılaştırılması

Araştırma projelerine ait anahtar kelimeler ile anahtar kelimelerin bağlantılı olduğu temel alan ve bilim alanlarından oluşan veri seti üzerinde K-Means, X-Means ve EM algoritmaları uygulanarak kümeleme analizi Weka uygulaması üzerinde gerçekleştirilmiştir. Elde edilen bulgular Tablo 4.6.'da yer almaktadır.

Tablo 4.6. K-Means, X-Means ve EM Kümeleme Algoritmaları Sonuçlarına ait Tablo

Algoritmalar	Min. Küme Değeri	Min. Küme Oranı- %	Max. Küme Değeri	Max. Küme Oranı-%	Hesaplama Süresi
K-Means	172	4	942	22	0.22
X-Means	160	4	1085	25	0.17
EM	56	1	865	20	0.98

K-Means ve X-Means algoritmalarının en küçük kümelerine ait değerlerin yakınlığı gözlemlenirken en büyük kümelerine sahip birim değerleri arasındaki değerlerin artışı Tablo 4.6.'da görünmektedir.

K-Means ve EM algoritmalarında en büyük kümelerin değerleri yakınlık gösterirken en küçük küme değerleri arasındaki fark Tablo 4.6.'da görüldüğü üzere yüksektir.

Kümeleme işlemi en hızlı 0.17 sn ile X-Means algoritmasında gerçekleşirken en yavaş 0.398 sn ile EM algoritmasında gerçekleşmiştir.

BÖLÜM 5. TARTIŞMA VE SONUÇ

Sakarya Üniversitesi Bilimsel Araştırma Proje Yönetim Sistemi 6 Mayıs 2018 tarihinden itibaren aktif olarak kullanılmaktadır. Araştırma projeleri için proje yürütücüleri tarafından başvurular bu platform üzerinden kayda alınmakta ve her proje için proje yürütücülerinin yapacağı çalışma alan bilgileri anahtar kelime olarak ilgili sistemin veri tabanında tutulmaktadır.

Projelerin çalışma alan bilgileri bağlı olduğu anahtar kelimeler, temel alan ve bilim alanları ile ilişkilidir.

Araştırma projeleri üzerinde yer alan YÖK bazlı 12 temel alanın tamamı yer almaktadır, yalnız proje yürütücüleri tarafından bu temel alanların 11 tanesi kullanılmıştır. Bunun sebebini ise kurumda Ziraat Fakültesinin olmamasının etkisidir.

Yapılan çalışmaların en büyük orana sahip kısmı %40.3 lük dilimi ile Mühendislik alanı üzerinde olduğu, yapılan çalışmalarda en az orana sahip alanın ise %0.3 lük dilim ile Hukuk alanına ait olduğu görülmektedir. Bunun sebebinin ise kurumun mühendislik temelli bir oluşuma sahip olması ve Hukuk Fakültesinin ise yeni açılan bir fakülte olması üzerine çalışma faaliyetlerinin henüz kısıtlı kalmasının etkisi olduğu düşünülebilir.

Araştırma projeleri üzerinde yer alan YÖK bazlı 12 temel alanın altında yer alan 300 bilim alanının 279 tanesi kullanılmıştır.

Araştırma projelerine ait toplamda ait 3551 adet anahtar kelime kullanılmıştır.

Araştırma projelerine ait veri seti üzerinde yer alan temel alan ve bilim alanı değerlerine uygulanan kümeleme analizleri algoritmaları neticesinde K-Means, X-Means ve EM algoritmaları içerisinde en hızlı kümeleme analizi X-Means algoritması ile 0.17 sn de gerçekleşmiştir.

Anahtar kelimeler üzerinde yer alan temel alan ve bilim alanlarının kümelenmesinde K-Means, X-Means ve EM algoritmaları uygulanmıştır ve en stabil değerlere X-Means algoritması ile ulaşıldığı gözlemlenmiştir.

Yapılan çalışmanın kurumların araştırma projelerini hangi doğrultuda gerçekleştirdikleri ve geleceğin yapılanması açısından yön göstergesi olabileceği düşünülmektedir. Kurumun geleceğini yapılandırırken çalışma alanları üzerinden ilgili model yapısı baz alınarak proje türlerinin oluşturulması ve gelecek yatırımlarını bu doğrultuda yapması hem kurumun vizyonunun göstergesi olacaktır hem de mali dağılımın kontrollü olarak aktarımına olananak sunacaktır.

KAYNAKLAR

Albayrak, S. (2019). Okul Servisi Araçlarını Rotalama Problemi İçin Yenilikçi Bir Yaklaşım (Master's thesis, Fen Bilimleri Enstitüsü).

Akat, Y. (2007). Ülkelerin Askeri Benzerliklerine Göre Kümeleme Analizi Yardımıyla Sınıflandırılması (Doctoral dissertation, Fen Bilimleri Enstitüsü).

Akaner, M. 2004. Gazi Üniversitesi Bilimsel Araştırma Projeleri Müdürlüğü'nde karar destek sisteminin oluşturulması. Gazi Üniversitesi, Fen Bilimleri Enstitüsü, Endüstri Mühendisliği Anabilim Dalı, Yüksek Lisans Tezi, Ankara.

Akın, Y. K. 2008. Veri madenciliğinde kümeleme algoritmaları ve kümeleme analizi.

Akkuş, B. 2017. Veri Madenciliği Yöntemleri ile Ülkeleri Gelişmişlik Ölçütlerine Göre Kümeleme Üzerine Bir Uygulama. İstanbul Aydın Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, Bilgisayar Mühendisliği Programı, İstanbul.

Aksakallı, Ö. 2019. Lenfoma Ameliyat Verilerini Kullanarak K-Means, Hiyerarşik ve EM Algoritmalarının Performanslarının Karşılaştırılması. Düzce Üniversitesi, Fen Bilimleri Enstitüsü, Elektrik- Elektronik ve Bilgisayar Mühendisliği (Disiplinlerarası) Anabilim Dalı, Yüksek Lisans Tezi, Düzce.

Anderberg, M.R. 1973. "Cluster Analysis for applications", Academic Press, 553-555, New York.

ARIBAŞ, M., & ÖZCAN, U. (2016). Akademik araştırma projelerinin AHP ve TOPSIS yöntemleri kullanılarak değerlendirilmesi. Politeknik Dergisi, 19(2), 163-173.

Arıcan, İrfan. 2010. Üniversite Sanayi İşbirliği Çerçevesinde Bilimsel Araştırma Projeleri ve Marmara Üniversitesi Araştırma Geliştirme Faaliyetlerinin Değerlendirilmesi. Marmara Üniversitesi, Sosyal Bilimler Enstitüsü, İşletme Anabilim Dalı, Yönetim ve Organizasyon Bilim Dalı, Yüksek Lisans Tezi, İstanbul.

ATBAŞ, A.C. 2008. Kümeleme analizinde küme sayısının belirlenmesi üzerine bir Çalışma. Ankara Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik Ana Bilim Dalı, Yüksek Lisans Tezi, Ankara.

Avcı, E., ÖZENİR, Ö. S., & YÜCEL, E. (2016). TÜBİTAK ortaöğretim öğrencileri araştırma projeleri yarışmasına katılan öğrencilerin yarışma sonrası kazanımlarının incelenmesi. Uşak Üniversitesi Sosyal Bilimler Dergisi, 9(27/3), 1-21.

Azak, E. N., & Şen, H. (2019). Bireylerde Organ Bağışını Etkileyen Faktörler: OECD Ülkeleri için Bir Araştırma. Ankara Hacı Bayram Veli Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 21(3), 535-547.

Bircan, H., & Çam, S. (2016). VERİ MADENCİLİĞİNDE KÜMELEME ANALİZİ VE SAĞLIK SEKTÖRÜNDE BİR UYGULAMASI. Cumhuriyet Üniversitesi İktisadi ve İdari Bilimler Dergisi, 17(2), 85-96.

Blashfield, R.K., & Aldenferder, M.S. 1978. " The Literature on Cluster Analysis", Multivariate Behavioral Research, 13, 271-295.

Boz, M. 2011. Üniversitelerde Bilimsel Araştırma Projelerinin (BAP) Etkinliği ve Mali Yapısının İncelenmesi: Çanakkale Onsekiz Mart Üniversitesi Örneği,. Çanakkale Onsekiz Mart Üniversitesi, Sosyal Bilimleri Enstitüsü, Maliye Anabilim Dalı, Yüksek Lisans Tezi, Çanakkale.

Burn, D. H., Zrinji, Z., & Kowalchuk, M. (1997). Regionalization of catchments for regional flood frequency analysis. Journal of Hydrologic Engineering, 2(2), 76-82.

Çakmak Zeki, 1999. "Kümeleme Analizinde Geçerlilik Problemi ve Kümeleme Sonuçlarının Değerlendirilmesi", Dumlupınar Üniversitesi Sosyal Bilimler Dergisi, Sayı:3, Kasım,s.187-205.

Çetintaş, H. (2019). TÜBİTAK ortaokul öğrencileri araştırma projelerinin bilimsel danışmanlık süreci yönetimi: Fen bilimleri örneği. Kütahya Dumlupınar Üniversitesi Eğitim Bilimleri Enstitüsü, Matematik ve Fen Bilimleri Eğitimi Anabilim Dalı, Fen Bilgisi Eğitimi Bilim Dalı, Yüksek Lisans Tezi, Kütahya.

Çokluk, Ö., Şekercioğlu, G. & Büyüköztürk, Ş. (2016). Sosyal Bilimler İçin Çok Değişkenli İstatistik SPSS ve LISREL Uygulamaları. Ankara: Pegem Akademi.

Dahl, T. ve Naes, T. (2004). Outlier and group detection in sensory panels using hierarchical cluster analysis with the procrustes distance. Food Quality and Preference, 15, 195–208.

Demiralay, M., & Çamurcu, Y. 2005 "Cure, Agnes ve K-means Algoritmalarındaki Kümeleme Yeteneklerinin Karşılaştırılması", İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 8(2): 1-18.

Dinler, M. 2014. Kümeleme Analizi Yöntemlerinin Hayvancılık Verilerinde Karşılaştırılmalı Olarak İncelenmesi. Bingöl Üniversitesi, Fen Bilimleri Enstitüsü, ZOOTEKNİ Anabilim Dalı, Yüksek Lisans Tezi, Bingöl.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). Cluster analysis. John Wiley & Sons.

FADHIL, M. A. F. 2019. Yarasa Algoritması ile Tıbbi Veri Setlerinin Kümelenmesi. Selçuk Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, Yüksek Lisans Tezi, Konya.

FIRAT S. Ü.,(1997) Kümeleme analizi istihdamın sektörel yapısı açısından Avrupa ülkelerinin karşılaştırılması, İ.Ü.Sosyal Bilimler Dergisi, 3, 50-59.

FIRAT, M., DİKBAŞ, F., KOÇ, A. C., & GÜNGÖR, M. (2012). K-ortalamalar yöntemi ile yıllık yağışların sınıflandırılması ve homojen bölgelerin belirlenmesi. Teknik Dergi, 23(113), 6037-6050.

Green, E.P. 1989. Analysing Multivariate Data, Philadelphia, p.427

GÜNAY ATBAŞ, A. C. Y. (2008). Kümeleme analizinde küme sayısının belirlenmesi üzerine bir çalısma (Doctoral dissertation, Ankara Üniversitesi Fen Bilimleri İstatistik Ana Bilim Dalı).

Gürbüz, M., & Karabulut, M. 2009. SSCB'nin dağılmasıyla bağımsızlığına kavuşan ülkelerde sosyo-ekonomik benzerlik analizi. Bilig Türk Dünyası Sosyal Bilimler Dergisi, 50, 31-50.

Gürel, A. G. 2019. Üniversite kütüphanesi verileri üzerinde veri madenciliği yöntemlerinin uygulanması. Afyon Kocatepe Üniversitesi, Fen Bilimleri Enstitüsü, İnternet ve Bilişim Teknolojileri Yönetimi Anabilim Dalı, Yüksek Lisans Tezi, Afyon.

Gökalp, S. 2014. Veri Madenciliğinde Çeşitli Kümeleme Algoritmalarının Farklı Platformlarda Karşılaştırmalı Analizi. Atatürk Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, Yüksek Lisans Tezi, Erzurum.

Hair, Jr. J. F., R. E. Anderson, R. L. Tatham, W. C. Black, Multivariate Data Analysis, Prentice-Hall, Upper Saddle River, New Jersey, 1998.

He, Q. (1996). A review of clustering algorithms as applied in IR. Graduate School of Library and Information Science University of Illinois at Urbana-Compaign. http://out.uclv.edu.cu/cei/

Hubert, L. (1974). Approximate evaluation techniques for the single-link and complete-link hierarchical clustering procedures. Journal of the American Statistical Association, 69(347), 698-704.

Işık, M. 2006. Bölünmeli Kümeleme Yöntemleri ile Veri Madenciliği Uygulamaları. Marmara Üniversitesi, Fen Bilimleri Enstitüsü, Elektronik Bilgisayar Eğitimi Anabilim Dalı, Bilgisayar ve Kontrol Eğitimi Programı, Yüksek Lisans Tezi, İstanbul.

Kaufman, L. & Rousseeuw, P.J. 1987. "Clustering by means of medoids", Statistical Data Analysis Based on The L1– Norm and Related Methods, pp. 405–416.

Kaufman, L. R., & Rousseeuw, P. PJ (1990) Finding groups in data: An introduction to cluster analysis. Hoboken NJ John Wiley & Sons Inc, 725.

Kaya, H. 2018. Akciğer Hastalıkları Teşhisinde Sınıflandırma ve Bulanık Mantık Yöntemlerinin Uygulanması. Ankara Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, Yüksek Lisans Tezi, Ankara.

Kılınç, O. 2019. Yatırım Planlamasında Meteorolojik Veriler ile Veri Madenciliği Kümeleme Uygulamaları. Milli Savunma Üniversitesi, Alparslan Savunma Bilimleri Enstitüsü Müdürlüğü, Harekat Araştırması Ana Bilim Dalı, Harekat Araştırması Programı, Yüksek Lisans Tezi, Ankara.

Lin, G. F., & Chen, L. H. (2006). Identification of homogeneous regions for regional frequency analysis using the self-organizing map. Journal of Hydrology, 324(1-4), 1-9.

Ulaş, A. 2016. Mikrodizi Verileri Üzerinde Kümeleme Algoritmalarının Uygulaması. Erciyes Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, Yüksek Lisans Tezi, Kayseri.

Üstünel, M. 2018. K - Ortalamalar Algoritmasına Dayalı Kümeleme Analizi Sistemi ve Perakendecilik Sektöründe Uygulaması. Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Matematik Mühendisliği Anabilim Dalı, Matematik Mühendisliği Programı, Yüksek Lisans Tezi, İstanbul.

Özdamar, K., 2004. Paket Programlar ile İstatistiksel Veri Analizi (Çok Değişkenli Analizler), Kaan Kitapevi, 502s. Eskişehir.

Resmi Gazete, 2020. 03.06.2020 tarihinde https://www.resmigazete.gov.tr/eskiler/2016/11/20161126-8.htm adresinden erişildi.

Şeker, S.E. 2020. 09.06.2020 tarihinde
http://bilgisayarkavramlari.sadievrenseker.com/2009/06/01/weka/ adresinden erişildi.

Saygılı, A. 2013. Veri Madenciliği ile Mühendislik Fakültesi Öğrencilerinin Okul Başarılarının Analizi. Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, Bilgisayar Mühendisliği Programı, Yüksek Lisans Tezi, İstanbul.

Sharma, S. 1996. "Applied Multivariate Techniques", Wiley-Interscience, New York

Savaş, S., Topaloğlu, N., & Yılmaz, M. 2012. Veri madenciliği ve Türkiye'deki uygulama örnekleri. İstanbul Ticaret Üniversitesi, Fen Bilimleri Dergisi Yıl:11 Sayı: 21 Bahar 2012 s. 1-23

Talan, M. İ. (2016). Veri Madenciliği ile Karpal Tünel Sendromuna Yönelik Ön Tanı Destek ve Hasta Takip Sisteminin Geliştirilmesi (Doctoral dissertation).

Tatlıdil, H., Uygulamalı Çok Değişkenli istatistiksel Analiz, Ziraat Matbaacılık A.Ş., Ankara, Eylül 2002

Tekingöz, Ö. 2016. Dijital Yayıncılıkta İçerik İzleme Oranlarına Göre Müşteri Kümelenmesi. Bahçeşehir Üniversitesi, Fen Bilimleri Enstitüsü, Bilgi Teknolojileri Anabilim Dalı, Yüksek Lisans Tezi, İstanbul.

ÖZGEÇMİŞ

Gülizar Pat, 05.01.1987'de İstanbul'da doğdu. İlk, orta ve lise eğitimini İstanbul'da tamamladı. 2005 yılında Selçuk Üniversitesi Bilgisayar Teknolojileri ve Programlama önlisans bölümünden mezun oldu. 2013 yılında Sakarya Üniversitesine Bilgisayar İşletmeni olarak atandı. Sakarya Üniversitesi Bilgisayar ve Bilişim Bilimleri Fakültesi Bilgisayar Mühendisliği Pr.'dan 2016 yılında mezun oldu. Yüksek lisans eğitimine 2018 yılında Sakarya Üniversitesi Fen Bilimleri Enstitüsü Bilişim Sistemleri Mühendisliği bölümünde başladı. Halen Sakarya Üniversitesinde Bilgisayar Araştırma ve Uygulama Merkezinde yazılım geliştirici olarak görev yapmaktadır.