Uluslararası İstatistiksel Sınıflamalara Yönelik Kod Atama Sistemi (KASİS)

Araştırma Makalesi/Research Article

Levent AHİ[1], Ebru KILIÇ ÇAKMAK[2]

[1] Türkiye İstatistik Kurumu, Ankara, Türkiye
[2] Bilgisayar ve Öğretim Teknolojileri Eğitimi Bölümü, Gazi Üniversitesi, Ankara, Türkiye
leventahi@tuik.gov.tr, ekilic@gazi.edu.tr
(Geliş/Received:07.07.2019; Kabul/Accepted:26.06.2020)
DOI: 10.17671/gazibtd.588097

Özet— İstatistiksel sınıflamaların, ülkelerin istatistik sistemlerinde çok büyük bir yeri ve önemi bulunmaktadır. İstatistiksel sınıflama kullanabilmenin yolu kod atamadan geçmektedir. Kod atama, elimizdeki metinsel tanım ile standart sınıflama sözlüğünde yer alan tanımı eşleştirme ve bu tanıma karşılık gelen sözlükteki kodu kullanma işleminden oluşmaktadır. Anketlerde, değişkenleri doğru gruplarda sınıflayabilmek için metinsel tanımlar sıklıkla kullanılmaktadır. Değişkenlerin sınıflamasının doğru olarak yapılmış olması bu değişkenler ile yapılacak araştırmaların sonuçlarının doğru olmasını sağlayacaktır. Kayıt sayısı arttıkça, değişkenlerin doğru gruplarda sınıflandığını kontrol etmek için manuel yöntemler yeterli olmayacaktır. Bu yüzden bu işlemi yapabilecek otomatik bir sisteme ihtiyaç duyulmaktadır. Bu çalışmada, sınıflama kullanan değişkenlerin doğru grupta sınıflanıp sınıflanmadığını otomatik şekilde kontrol edebilen sistem tanıtılmaktadır. Sistemin etkinliği, Türkiye İstatistik Kurumu'nun (TÜİK) yapmış olduğu Hanehalkı Bütçe Araştırması (HBA) 2017 yılı veri seti kullanılarak değerlendirilmiştir. Bu veri seti, ülkemizdeki tüketim harcamaları istatistiklerinin ana kaynağıdır. Tüketim harcamalarının sınıflamasında Uluslararası Bireysel Tüketimin Amaca Göre Sınıflaması (COICOP) kullanılmaktadır. Anketör tarafından kod ataması yapılmış kayıtlar, geliştirilen sistem ile kontrol edilerek sonuçları incelenmiştir. Geliştirilen sistem, denetimli makine öğrenmesi yöntemlerini kullanan sistemlerden eğitim veri kümesine ihtiyaç duymaması ile ayrılmaktadır. Sıfır noktasından itibaren sistem çalışmaya başlayabilir ve her bir ilave kayıtta kendi öğrenmesini artırarak devam etmektedir. Bu sistem, eğitim veri kümesindeki kayıtların sınıflamasının doğru olarak yapılıp yapılmadığını kontrol ederek denetimli makine öğrenmesi yöntemini kullanan sistemlerin doğru şekilde öğrenmelerine de katkı sağlayabilmektedir.

Anahtar Kelimeler— istatistiksel sınıflama, COICOP, HBA, otomatik kodlama, metinsel tanım

Code Assignment System (KASIS) for International Statistical Classifications

Abstract— Statistical classifications have a great place and importance in the statistical systems of countries. The way to use statistical classification is code assignment. Code assignment consists of matching the textual definition with the definition in the standard classification dictionary and using the code in the dictionary corresponding to this definition. In questionnaires, textual definitions are often used to classify variables in right groups. The correct classification of the variables will ensure that the results of the studies to be conducted with these variables are correct. As the number of records increases, manual methods will not be sufficient to check that variables are classified in the correct groups. Therefore, there is a need for an automated system that can perform this process. This study introduces a system that can automatically check whether the variables using classification are classified in the correct group. The effectiveness of the system was tested using the 2017 Household Budget Survey (HBS) micro data set made by Turkey Statistical Institute (TURKSTAT). This data set is the main source of consumption expenditure statistics in our country. Classification of Individual Consumption by Purpose (COICOP) is used in the classification of consumption expenditures. The code assignment made by the interviewer was checked with the developed system and the results were examined. The developed system differs from systems using supervised machine learning methods by not needing a training data set. Starting from the zero point, the system can start working and continues its learning by increasing its learning in each additional record. This system can also contribute to the correct learning of systems using this method by controlling whether the classification of records in the education data set has been made correctly or not.

Keywords— statistical classification, COICOP, HBS, automatic coding, textual definition

1. GİRİŞ (INTRODUCTION)

Küreselleşme ile birlikte birçok alanda kapsamlı bir dönüşüm süreci yaşanmaktadır. Bu dönüşüm sürecinde insan ihtiyaçları da değişmekte ve farklılaşmaktadır. İstatistik sistemleri de bu dönüşüme ayak uydurarak farklılaşan insan ihtiyaçlarına cevap verebilecek şekilde kendilerini güncellemek zorundadırlar [1].

Bu dönüşüm sürecinde, veriden anlamlı bilgiler çıkararak karar alıcılara sunabilmenin yolu istatistiksel çalışmalardan geçmektedir [2]. Çevremizde olup biten gelişmeleri takip edebilmek, yorumlayabilmek ve sonucunda gereken stratejik kararları alabilmek için güvenilir istatistiki bilgiye ihtiyaç duyulmaktadır.

Farklı coğrafyada yaşayan ve farklı kültüre sahip insanların aynı bilgiden aynı fayda veya sonucu elde edebilmeleri için birtakım düzenlemelerin yapılması gerekmektedir. İstatistiki sınıflamaların bu düzenlemeleri sağlamaları bakımından önemli görevleri bulunmaktadır [3, 4].

Bu kapsamda oluşturulmuş çok sayıda sınıflama türü bulunmaktadır. Faaliyet sınıflaması, ürün sınıflaması, amaca göre sınıflama bu sınıflama türlerinden bazılarıdır [5].

2. SINIFLAMALAR (CLASSIFICATIONS)

Sınıflama, birbirine benzeyen nesne, gözlem ve olayları belli bir amaca göre gruplara ayırma işlemidir [6]. Sınıflama işlemi sonucunda aynı grupta yer alan nesne, gözlem veya olaylar benzer özelliklere sahip olmaktadır.

İstatistikler, genellikle çok sayıda gözlemden hesaplama yapılarak üretilmektedir. Eğer gerekiyorsa, gözlemlerin benzerliklerine göre sınıflanması hesaplamaya başlamadan önce yapılması gereken bir faaliyettir. Bu nedenle, sınıflama faaliyeti istatistiklerin üretim aşamasında önemli bir adımdır ve karşılaştırılabilir veri üretebilmek için standart olarak oluşturulmuş sınıflamalar ülkelerin resmi istatistikleri için kilit araçlar olmaktadır [7].

Sınıflamalar konusu kendisine 5429 sayılı Türkiye İstatistik Kanunu'nda da yer bulmuştur. Bu kanunun "Sınıflamalar" başlıklı 11. maddesinde "Kurum ve kuruluşlar, Başkanlık tarafından oluşturulan istatistik amaçlı tanım ve sınıflamaları kullanmakla; kendi ihtiyaçları doğrultusunda belirleyecekleri sınıflamaları kullanmaları halinde ise, Başkanlığın belirlediği sınıflamalara dönüştürülmesini sağlayacak tedbirleri almakla yükümlüdür." denilmektedir [8]. Üretilen istatistiklerin karşılaştırılabilir olması bakımından sınıflamaların önemli bir işlevinin olduğu bu kanun maddesinden anlaşılmaktadır.

Sınıflamalar, alanlarında uzmanlaşmış Birleşmiş Milletler İstatistik Bölümü (UNSD), Avrupa Birliği İstatistik Ofisi (Eurostat), Uluslararası Çalışma Örgütü (ILO), Uluslararası Para Fonu (IMF), İktisadi İşbirliği ve Kalkınma Teşkilatı (OECD), Birleşmiş Milletler Eğitim, Bilim ve Kültür Örgütü (UNESCO) gibi uluslararası kuruluşlar tarafından geliştirilmektedir [5].

Ülkeler kendi ihtiyaçlarını karşılamak amacı ile ulusal sınıflamalarını oluşturarak kullanabilmektedir. Bu durumda, ülkeler arasında karşılaştırılabilir bir veri üretilmesi söz konusu ise ulusal sınıflamadan uluslararası bir sınıflamaya dönüşüm sağlayacak anahtarların kullanılması gündeme gelmektedir [5]. Bu nedenle, referans olarak kullanılabilecek uluslararası sınıflamaların olması ulusal istatistik ofislerine kolaylıklar sağlamaktadır.

2.1. Kodlama ve Kodlama Sistemleri (Coding ve Coding Systems)

İstatistikler, genel olarak anket yapılarak üretilmektedir. Anketlerde, cevabı metinsel olan sorular yöneltilmekten mümkün olduğunca kaçınılmaktadır. Ancak, anketlerde kapalı sorular sormak her zaman mümkün olmamaktadır [9]. Bu yüzden her araştırmada metinsel cevapları doğru şekilde sınıflamak ve kodlamak kaçınılmaz bir iş olmaktadır. Kodlama, bir veya birkaç kelimeden oluşan metinsel bir ifadeyi bir terminoloji ile ilgili olan sınıflamaya ait bir koda dönüştürme anlamına gelmektedir.

Anketlerde, sınıflama gerektiren sorular ile ilgili cevaplar metinsel olarak alınarak yanlış sınıf seçimi engellenmeye çalışılmaktadır. Metinsel cevaplarda yapılan hatalar, anket sonuçları üzerinde daha az etkili olmaktadır. Aynı zamanda, kendisine anket uygulanan kişi bu tür sorulara kendi kelimeleri ile cevap verebilme özgürlüğüne sahip olmaktadır [10]. İhtiyaç duyulan formda ve sınıflamada bir cevap verilmesi istenmesi durumunda, cevap verenlerin iktisadi faaliyet, meslek, eğitim vb. sınıflamalar hakkında bilgi sahibi olması gerekmektedir. Ancak, sınıflama konusunda uzman olmayan birinden böyle bir şey beklemek akla yatkın değildir.

Sınıflama probleminin çözümü, son yıllarda makine öğrenmesi ve veri madenciliğinin önemli çalışma alanlarından biri olmuştur [11].

Kodlama gerektiren çok sayıda sınıflama bulunmaktadır. Hacking ve Willenborg [10], Hollanda İstatistik Ofisi'nde kullanılan ve kodlama gerektiren sınıflamaları, eğitim, meslek, ürün, endüstri ve ölüm nedenleri olarak sıralamaktadır.

Groves ve diğ. [12], kodlamanın gerekli olduğu sınıflamaları Standart Meslek Sınıflaması, Kuzey Amerika Endüstri Sınıflama Sistemi ve Uluslararası Hastalık Sınıflaması olarak vermektedir.

Esuli ve Sebastiani [13] pazar araştırması problemine kodlama uygulamışlardır. Bu kapsamda, bir ürün veya marka adı ile cevap verilen "favori meşrubatınız nedir?" şeklinde bir soruya verilen cevapları sınıflamışlardır.

Kodlama faaliyetini yerine getirmek için farklı yöntemler mevcuttur: manuel kodlama, bilgisayar destekli kodlama ve otomatik kodlama. Kodlama faaliyeti için hangi

seçeneğin en iyi olduğu sınıflama ve verinin karmaşıklığına bağlı olmakla birlikte bu sistemlerin kombinasyonları da uygulamada kullanılabilmektedir [9].

2.1.1. Manuel Kodlama (Manuel Coding)

Metinsel cevapların kodlanması birkaç yıl öncesine kadar manuel olarak yapılmaktaydı. Manuel yöntemlerle yapılan bu iş çok zaman alıcı ve maliyetliydi. Aynı zamanda, sürecin standardizasyon ve atanmış kodun doğruluğunun sağlanması konusunda da bir garantisi bulunmamaktaydı. Kanada İstatistik Ofisi tarafından geliştirilen otomatik kodlama programının 1991 Kanada Nüfus Sayımı'nda dört aylık kullanımı sonucunda, kodlamadaki doğruluk oranı artmış, kodlayıcı sayısı 600'den 25'e inmiş ve 5,9 milyon ABD doları bütçeye sahip projede 3,5 milyon dolardan fazla tasarruf sağlanmıştır [14].

İrlanda örneğinde, manuel meslek kodlamasının, çok sayıda kodlama personeline ihtiyaç duyması, yüksek maliyetler gerektirmesi, kodlamanın kalitesi veya doğruluğunun yeterince tatmin edici olmaması ve aynı tanıma ilişkin verilen kodun, farklı kodlayıcılar arasında farklılık göstermesi dezavantajlarına sahip olduğu belirlenmiştir [15].

2.1.2. Bilgisayar Destekli Kodlama (Computer-assisted Coding)

Kodlamadaki etkinliği artırmak amacıyla zaman içerisinde daha yaygın olarak bilgisayar kullanılmaya başlanmıştır [10]. Türkiye İstatistik Kurumu (TÜİK) yaptığı anketlerde bilgisayar destekli kodlama uygulaması kullanmaktadır.

Bilgisayar destekli kodlama, kodlayıcı tarafından gözlem bazında gerçekleştirilen ve kodlama görevini kolaylaştıran tüm teknikleri içermektedir. Bu yaklaşım, bazen otomatik bir kodlama sisteminden daha uygun bir kod bulma şansı vermekle birlikte kodlama yapacak kişinin sınıflama hakkında üst düzey bilgisini gerektirmektedir.

Anketlerde, cevap veren tarafından kodlama bazı sorularda örtülü olarak yapılmaktadır. Yani, verilen cevaba göre en uygun seçenek işaretlenmektedir.

Bazı durumlarda, cevap verene açık uçlu bir soru sorulmakta ve cevap profesyonel bir kodlayıcı tarafından kategorize edilmektedir. Bu yöntemde, kodlayıcı işini yaparken cevap veren karşısında olmayacağı için bazı detaylar eksik kalabilmektedir. Bu durumda, görüşme sırasında gerekli tüm bilgileri sorgulayabilen anketöre kodlama görevi verilmesi önerilmektedir [16].

Bilgisayar destekli kodlamada, özel olarak tasarlanmış bilgisayar programları kodlama görevini kolaylaştırmak için kullanılmaktadır. Hangi kodun doğru olduğuna kodlayıcı karar vermektedir. Program genellikle az sayıda kod önererek kodlayıcıya yardım etmektedir.

Meslek kodları için Bushnell [17], bir bilgisayar programının kodlama sürecini hızlandırabileceğini ve aynı zamanda kodlama kalitesini artırabileceğini göstermiştir.

Açık uçlu soruların anketlere entegrasyonu uzun süredir devam eden bir metodolojik kaygıdır. Bu sorunu giderebilmek amacıyla çeşitli program ve yazılım çözümleri tanıtılmıştır. Fielding ve diğ. [18] açık uçlu sorularına verilen cevapların analiz edilmesi gerektiğini belirtmişler ve Bilgisayar Destekli Nitel Veri Analizi (CAQDAS) yazılımının böyle bir bağlamda faydalı bir kaynak olabileceğinden bahsetmiştir. Yazılım, açık uçlu sorulara verilen cevapların analizine ilgi duyan araştırmacılara yöneliktir.

Esuli ve Sebastiani [13] kullanıcı tarafından belirlenen herhangi bir kod çerçevesinde metinsel cevapları otomatik olarak kodlayan ve kullanımı son derece pratik olan Kodlama Sistemi (VSC)'ni tanıtmıştır.

Hollanda İstatistik Ofisi, bilgisayar destekli kodlamada kullanılan yeni bir yöntem geliştirmişlerdir. Bu yeni yönteme göre, kod ataması yapmak için girilen metin üç harflik kısımlara ayrılarak sözlükte arama gerçekleştiriliyor. Uygun sonuçlar kodlayıcıya gösterilerek içerisinden uygun olanı seçmesi sağlanmaktadır [19].

2.1.3. Otomatik Kodlama (Automatic Coding)

Otomatik kodlama, herhangi bir müdahale gerektirmeyen bir kodlama algoritmasının otomatik olarak uygulanması anlamına gelmektedir.

Clarke ve Brooker [20], kod atama işleminin doğrudan insan katılımı olmadan bilgisayar tarafından yapılmasını otomatik kodlama olarak tanımlamaktadır.

Otomatik kodlamada genellikle bir listedeki kayıtlara kod atama işlemi yerine getirilmektedir. Bu türden bir algoritma başarılı bir şekilde çalıştığında, atanan kod daima tek bir kod olmaktadır. Otomatik kodlamanın kalitesi, doğruluğu ve hızı ile ölçülebilmektedir [10].

Otomatik kodlamada ele alınması gereken iki büyük sorun bulunmaktadır [10]:

1. Yazılı metnin yorumlanması ve
2. Tanımların uyması gereken karmaşık sınıflama.

Anketör tarafından yazılan metinlerle ilgili olarak aşağıdaki birçok sorun ortaya çıkabilmektedir. Ancak bu sorunlara rağmen, bilgisayar programı metne en uygun kodu sözlükten seçebilmelidir [10].

1. Yazım sorunları
2. Dil bilgisi problemleri (kelimeler arasındaki ilişkiler, söz dizilimleri)
3. Anlamsal problemler (kelimelerin anlamı, kavramlar, cümle parçaları, tek bir cümle, birkaç cümle)
4. Yorumlama problemleri (sınıflamadaki hangi kod metne en iyi şekilde uyar).

Metinle ilgili ortaya çıkabilecek bir başka sorunda, metnin kodlama yapmaya yetecek kadar ayrıntı içermemesi veya metnin sözlükteki iki veya daha fazla koda karşılık

gelmesidir. Bu sorun, metnin gerekenden daha ayrıntılı ifadeler içermesi nedeniyle de ortaya çıkabilmektedir [10].

Metinsel bir cevabı otomatik olarak kodlamak için bir sözlüğe veya daha önce kodlanmış diğer kayıtlara ihtiyaç duyulmaktadır [9]. Bu kayıtlar içerisinde yapılan arama işlemi, doğrudan insan katılımı olmadan bir bilgisayar tarafından yapıldığında, otomatik kodlama olarak adlandırılmaktadır.

Otomatik kodlama, basit olarak mantıksal kurallar kullanılarak da yapılabilmektedir. Bu durumda, tam olarak belirtilen koşul gerçekleşince öngörülen kod atanmaktadır. Meslek kodlaması için, çeşitli yazarlar bu tekniği uygulamışlardır. Geis [21], Drasch ve diğ. [22], Jung ve diğ. [23] ve Conrad [24] bu tekniği kodlamanın ilk adımı olarak kullanmışlardır. Bu sistemlerde 1000 tane kural olsa dahi, meslek kodlarının doğru bir şekilde kodlanmasının garantisi bulunmamaktadır. Kuralda yer almayan yeni bir kayıt geldiğinde sistem bir sonuç üretemeyecektir. Bu yüzden, Hartmann ve Schütz [25] kodlamada daha yüksek başarı oranları için daha fazla kural oluşturmuşlardır.

Otomatik kodlama uygulamalarında son yıllarda denetimli makine öğrenmesi teknikleri kullanılmaktadır. Makine öğrenmesi, istatistiki verilerin işlenmesi veya işlenmesine yardımcı olmak için algoritmalar aracılığıyla mevcut bilgilerin kullanıldığı yapay zekanın bir uygulaması olarak tanımlanabilir [26].

Sınıflama problemlerinin çözümünde kullanılan makine öğrenimi algoritmasının seçiminde dikkat edilmesi gereken kriterlerden biri, algoritmanın genele uygulanabilme performansıdır. Bu performans ne kadar iyi olursa algoritmanın başarısı o denli iyi olacaktır.

Bethmann ve diğ. [27], Alman panel araştırmalarında otomatik meslek kodlaması için iki tür olasılıklı denetimli makine öğrenme algoritması uygulamış ve eğitim verisi olarak yaklaşık 300000 adet manuel kodlanmış meslek kodu ve tanımlarını kullanmışlardır. Yazarlar, algoritmanın girdi bilgisi olarak kullanılan eğitim verisinin kaliteli olması durumunda, meslek kodlarının otomatik olarak yüksek başarı ile kodlanabileceği sonucuna varmışlardır.

Belloni ve diğ. [28], meslek kodlarındaki hataları incelemek amacıyla Hollanda'daki "Avrupa'da Sağlık, Yaşlanma ve Emeklilik Anketi (SHARE)" verilerindeki son ve şu anki mesleğe ilişkin tanımları bir yazılım programı kullanarak yeniden kodlamışlardır. Daha önce yapılan kodlamada yüksek miktarda hata tespit etmişlerdir. Anketlerde, kodlama kalitesinin önemli olduğunu ve genellikle ihmal edildiğini belirtmişlerdir. Yanlış kodlamalar nedeniyle hataların istatistiksel analizler yapılırken veya ekonometrik modellerde dikkate alınması gerektiğine vurgu yapmışlardır.

İrlanda Merkez İstatistik Ofisi, daha önce kodlanmış kayıtları eğitim verisi olarak kullanarak, Bireysel Tüketimin Amaca Göre Sınıflaması (COICOP) için otomatik bir kodlama sistemi üzerinde çalışmaktadırlar [26].

Lüksemburg İstatistik Ofisi, Tüketici Fiyat Endeksi (TÜFE)'nin derlenmesinde kullanılan perakende satış işlemlerindeki ürün tanımlarına COICOP kodlarını atamaktadır. Proje, maliyetleri artırmadan daha büyük miktarda veriye kodlama yapılmasına izin vermektedir [29].

Kanada İstatistik Ofisi, ödeme işlemlerine COICOP kodu atamak için makine öğrenmesi yöntemlerini kullanmak için proje yürütmektedir. Projenin amacı; perakende satış, hanehalkı tüketimi, dijital işlemler ve turizm istatistikleri konusunda ödeme işlemleri verisinin kullanılmasını sağlamaktır [29].

Yeni Zelanda İstatistik Ofisi, sayımlardaki kodlama kalitesini artırmak için destek vektörü makineleri (SVM) algoritmasının kullanılma potansiyelini araştırmaktadır. Census 2013 verilerindeki meslek ve okul sonrası yeterlilik değişkenlerini kodlamak için SVM algoritmasını kullanmışlardır. Her iki değişken için test verilerinde elde ettikleri düşük başarı oranları neticesinde SVM algoritmasını otomatik kodlamada kullanabilmek amacıyla daha fazla araştırma yapılması gerektiğine karar vermişlerdir [26].

De Matteis ve diğ. [30], fertlerin geçmişte yaptıkları mesleklerin hızlı bir şekilde toplanmasını ve kodlanmasını sağlamak için yenilikçi bir web tabanlı araç geliştirmişlerdir.

Haslinger [31], Avusturya Nüfus Sayımı'nda yer alan çalışılan yer değişkeni için otomatik kodlama yapabilmek için n-grams tekniğini kullanmışlardır.

Literatürdeki örneklerden anlaşılacağı üzere, kodlamanın istatistiksel süreçte kullanılan bir faaliyet olduğu açıkça görülebilmektedir. Kodlama, doktorun kendisine çeşitli şikayetler ve semptomlar sunan hastaya tanı koymasına çok benzemektedir. Doktorun görevi, birkaç gözlem, hastadan gelen yanıtlar ve muhtemelen ek testlere (örneğin kan testleri) dayanarak teşhis koymak ve bu temelde bir tedavi seçmektir [32], [33], [34]. Benzer yaklaşımla, kodlama işlemi eldeki bilgiler ışığında metinsel tanıma en uygun kodu atama işlemidir.

Kodlama, metinsel bir cevabın önceden tanımlanmış olası metinlere uygunluğuna yönelik seçim kararının alındığı bir süreçtir. Bu seçim, bazen anket esnasında bazen daha sonra kodlayıcılar tarafından da yapılabilmektedir. Ancak bu aşamada cevap veren kişi genellikle bulunmaz ve ilave başka sorulara cevap veremez. Bu süreçte, cevaplayıcıların kendi kelimeleri ile verdiği cevaplar girdi, çıktı ise belirli bir sınıflamayla ilgili bir kod olmaktadır [10].

Tablo 1'de kodlamanın kim tarafından nerede yapıldığı ve yapılan kodlamanın avantaj – dezavantajları yer almaktadır.

Tablo 1. Kodlamanın olası yerleri ve kim tarafından / ne yapıldığı [10]?
(Possible places to code and by whom/what?)

Kim / Ne?	**Nerede?**	**Avantajları**	**Dezavantajları**
Cevaplayıcı	Bilgisayar destekli web görüşmesi (CAWI)	Doğrudan geribildirim.	Cevaplayıcının sınıflama bilgisinin olmaması.
Anketör	Bilgisayar destekli telefon görüşmesi (CATI), Bilgisayar destekli yüzyüze görüşme (CAPI), Kağıt ve kalem görüşmesi (PAPI)	Doğrudan geribildirim.	Anketörün sınıflama bilgisinin yüzeysel olması.
Kodlayıcı	Ofis	Sınıflama uzmanının varlığı, Ek bilgilerinin de kullanılması, Cevapların bir bilgisayar programından daha iyi yorumlanması.	Doğrudan geribildirim her zaman mümkün değildir, Geribildirim çok zaman alır, Kodlama tutarsız olabilir.
Bilgisayar programı	Ofis	Hızlı ve tutarlı kodlama, Kodlama bilgisi bir sistemde belirtildiği için transfer edilebilir, Toplu halde kodlama yapılır, Gece veya gündüz çalışabilir.	Doğrudan geri bildirim yok, Geribildirim çok zaman alır.

İstatistiksel araştırmalarla genellikle kategorize edilmesi gereken çok büyük miktarlarda veri üretilmektedir. Bu bakımdan verinin doğru sınıflanıp sınıflanmadığının kontrolünü yapmak bu sınıflamalar sonucunda bir analiz ve çalışma yapılacaksa büyük önem arz etmektedir. TÜİK alanda kodlama yaparken bilgisayar destekli kodlama yöntemini uygulamakla birlikte kodlaması yapılmış kayıtlar uzman personel tarafından gözle manuel olarak kontrol edilmektedir. Bunun için gerek daha önce kodlaması yapılmamış kodlamaları otomatik şekilde yapabilecek gerekse daha önce kodlaması yapılmış olan kayıtların doğru sınıfta sınıflanıp sınıflanmadığını kontrol edebilecek bir sisteme ihtiyaç duyulmaktadır.

Bu çalışmanın amacı ve temel motivasyon kaynağı, kodlama ve sınıflamada yaşanan sorunların tamamına çözüm olacak bir bilgisayar programı için gerekli altyapının oluşturulması, geliştirilmesi ve bu kapsamda bilgisayar programının tasarımını, geliştirmesini ve değerlendirmesini yapmaktır.

Ulusal ve uluslararası alanda standart bir sınıflamanın kullanılmasının yanında günümüzde her istatistiğin otomasyon ile insandan bağımsız otomatik yöntemlerle yapılması önemlidir. Geliştirilen sistem standart sınıflama sözlüğü olan tüm sınıflamalar için kullanılabilir olmakla birlikte çalışma sadece COICOP sınıflaması üzerinedir. Bu kapsamda, çalışmada 2017 yılında TÜİK tarafından uygulanan Hanehalkı Bütçe Araştırması (HBA) mikro veri setinde bulunan anketör tarafından ataması yapılmış tüketim harcamaları kodları kullanılmıştır.

3. HANEHALKI BÜTÇE ARAŞTIRMASI (HOUSEHOLD BUDGET SURVEY)

HBA, hanelerin tüketim harcamalarının seviye ve bileşimleri hakkında bilgi veren en önemli kaynaklardan biridir [35].

HBA ile;

- Tüketici fiyat endekslerinde kullanılacak maddeler ile ağırlıkların belirlenmesi,
- Hanelerin tüketim harcamalarındaki değişikliklerin izlenmesi,
- Milli gelir hesaplamalarında kullanılacak yardımcı değişkenlerin elde edilmesi,
- Mutlak yoksulluk sınırının belirlenmesi amaçlanmaktadır [35].

HBA kapsamında 2017 yılında, her ay 1296, yıllık 15552 hane ile anket yapılmıştır. Hanelerden bilgiler; görüşme, kayıt ve gözlem metotları kullanılarak derlenmiştir. Anketörler, haneleri ayda 6 kez ziyaret etmişlerdir.

Anket ayı başlamadan önce seçilmiş haneye yapılan ziyarette, hanelerin fert bileşimi, konut özellikleri, alışkanlıklarına ilişkin bilgiler alınmıştır. Anket ayındaki ziyaretlerde, hanenin yaptığı tüketim harcamaları kayıt altına alınmıştır. Anket ayı sonunda yapılan görüşmede, hanehalkı fertlerinin gelir bilgileri derlenmiştir. 2017 Hanehalkı Bütçe Anketi'nde cevapsızlık oranı %21,8 olarak gerçekleşmiştir [36].

İstatistik ofisleri dışında HBA verilerinin geniş bir kullanıcı yelpazesi bulunmaktadır [37].

1. Ekonomik ve sosyal politika planlama amaçları için kullanan bakanlık ve kamu kurumları,
2. Hanehalklarının yaşam koşulları üzerine araştırma yapan üniversiteler ve araştırma kuruluşları,
3. Özel firmalar ve danışmanlar,
4. Genel olarak kitle iletişim araçları aracılığıyla bilgi alan veya istatistik ofislerinin yayınlarını kullanan diğer kullanıcılar.

Bu kadar geniş bir kullanım alanı bulunan anket verilerindeki doğru sınıflanmış her bir değişkenin çok büyük önemi bulunmaktadır.

3.1. Tüketim Harcamaları Sınıflaması (Consumption Expenditures Classification)

HBA'da, tüketim harcamaları sınıflaması olarak COICOP kullanılmaktadır. COICOP, hanehalkı tüketim harcamalarının uluslararası sınıflamasıdır. Bu sınıflamanın amacı, hanehalkı tüketim harcamalarında yer alan mal ve hizmet kategorileri için bir çerçeve sağlamaktır. COICOP, Ulusal Hesaplar Sistemi'nin (SNA) ayrılmaz bir parçasıdır. Bununla birlikte, tüketici fiyat endeksleri, gayri safi yurtiçi hasıla (GSYİH), kültür, spor, yiyecek, sağlık ve turizm ile ilgili istatistiklerde de kullanılmaktadır [38].

Hanehalkı tüketim harcamalarının uluslararası düzeyde sınıflandırılması fikri, Ekim 1923'te Uluslararası Çalışma Örgütü (ILO) tarafından düzenlenen Uluslararası Çalışma İstatistiği Konferansı'nda dile getirilmiştir.

COICOP adı altında ilk sınıflama, Mart 1999'da SNA'nın diğer üç işlevsel sınıflandırmasıyla birlikte Birleşmiş Milletler İstatistik Komisyonu tarafından kabul edilmiştir. O günden sonra da birkaç kez revize edilerek bugünkü halini almıştır. Son olarak, Mart 2018'de revize edilmiştir [38].

Avrupa Bireysel Tüketimin Amaca Göre Sınıflaması (ECOICOP), COICOP sınıflamasının özel bir şeklidir. Orijinal ECOICOP sınıflaması beş seviyeden oluşan hiyerarşik bir yapıya sahiptir. 12 adet Düzey 1 kategorisi mevcuttur. Düzey 1 kategorileri aşağıda verilmiştir [5].

1. Gıda ve alkolsüz içecekler
2. Alkollü içecek, sigara ve tütün
3. Giyim ve ayakkabı
4. Konut ve kira
5. Ev eşyası
6. Sağlık
7. Ulaştırma
8. Haberleşme
9. Kültür, eğlence
10. Eğitim hizmetleri
11. Otel, lokanta ve pastane
12. Çeşitli mal ve hizmetler

47 adet Düzey 2, 117 adet Düzey 3 ve 303 adet Düzey 4 kategorisi bulunmaktadır. Örnek hiyerarşik yapı aşağıda Tablo 2'de verilmiştir [5].

Tablo 2. COICOP hiyerarşik yapısı örneği
(COICOP hierarchical structure example)

Kod	**Tanım**	
01	Gıda ve alkolsüz içecekler	
01.1	Gıda	
01.1.6	Meyveler	
01.1.6.1	Taze veya soğutulmuş meyveler	*Kapsananlar:* - karpuz ve kavun, üzümsü meyveler
01.1.6.2	Dondurulmuş meyveler	*Kapsananlar:* - dondurulmuş meyveler - dondurulmuş üzümsü meyveler
01.1.6.3	Kurutulmuş meyveler ve sert kabuklu meyveler	*Kapsananlar:* - kurutulmuş meyveler, meyve kabukları, meyve çekirdekleri - sert kabuklu meyveler ve yenilebilir tohumlar - kuru üzümsü meyveler
01.1.6.4	Korunmuş meyveler ve meyve bazlı ürünler	*Kapsananlar:* - korunmuş meyveler ve meyve bazlı ürünler - diyet müstahzarları ve yemeklere tat vermek için kullanılan yalnızca meyve bazlı malzemeler - konserve edilmiş veya kutulanmış meyveler

Avrupa Birliği ülkelerinde HBA anketleri centilmenlik sözleşmesi kapsamında yürütülmektedir. Bu anlaşmaya 1989 yılında İstatistik Programı Komitesi toplantısında ulaşılmıştır [39]. Eurostat, üye ülkeler ve aday ülkelerden HBA veri setlerini beş senede bir beş basamaklı Düzey 4 kodları bazında tüketim harcamalarını içerecek şekilde göndermelerini istemektedir. Anket, centilmenlik anlaşmasına dayanarak yapıldığından, her ülke anketin amacı, metodolojisi ve çalışma sıklığına kendisi karar vermektedir. Anket uyumu konusunda çaba sarf edilse bile farklılıklar halen devam etmektedir. Ülkeler arasında HBA anketleri bakımından sıklık, zamanlama, içerik veya yapı bakımından farklılıklar bulunmaktadır.

TÜİK, HBA anketinde tüketim harcamaları sınıflaması olarak 10 basamaklı ECOICOP kodlarını kullanmaktadır ve her beş senede bir beş basamaklı Düzey 4 kod ayrımında tüketim harcama veri setlerini Eurostat'a göndermektedir.

4. MATERYAL VE METOT (MATERIAL AND METHOD)

HBA'da, seçilmiş hanelerin tüketim harcamaları veri giriş programı aracılığıyla anketörler tarafından veri tabanına kaydedilmektedir. Bu kapsamda veri giriş programında dört farklı tablo türü bulunmaktadır.

1. **Dayanıklı tablosu:** Hanelerin son 11 ay içinde satın aldığı ve sıklıkla alınması beklenmeyen seyrek düzeyde yapılan harcama kayıtlarını kapsamaktadır.

2. **Stok tablosu:** Hanelerin anket ayından önce satın aldıkları stok kayıtlarını kapsamaktadır.

3. **Satınalış tablosu:** Hanelerin anket ayında satın aldığı mal ve hizmet kayıtlarını kapsamaktadır.

4. **Ayni tablosu:** Hanelerin anket ayı içinde satın aldıkları ayni mal ve hizmet kayıtlarını kapsamaktadır.

Örnek olarak hanenin makarna için harcama yaptığını varsayalım. Şekil 1'de anketör veri giriş programı ekran görüntüsü verilmiştir. Anketör, Şekil 1'e göre 76. sırada harcamanın adı, cinsi ve tanımı alanına "makarna" yazmıştır. Kod alanında, yazılan tanıma uygun olarak içerisinde makarna ifadesi geçen harcama tanımları ve kodları Şekil 2'de verilmiştir.

Şekil 2'ye göre, içerisinde makarna ifadesi geçen 10 farklı alternatif kod bulunmaktadır. Şekil 3'te bu 10 farklı alternatif koddan hanenin yaptığı harcamaya uygun olan "111601030- Makarna Spagetti (Sade)" kodu anketör tarafından seçilmiştir. Harcamaya ilişkin miktar, ölçü birimi ve toplam değer girilerek bu harcama için veri giriş işlemi tamamlanmış olmaktadır.

Defterdeki sıra no	Harcamanın adı, cinsi ve tanımı	Markası	Kodu	Miktarı	Ölçü birimi	Toplam değeri	
71	ekmek normal		111301010	0,6	4470	3	5,0
72	bira teneke kutu alkollü 50 cl	efes	213101020	2	2600	18	9,0
73	lokantada yenen tabldot yem		1111118010	1	4520	8	8,0
74	yaş pasta		111401010	1	2600	50	50,0
75	apartman ortak harcamaları		444101010	1	4560	50	50,0
76	makarna						

Şekil 1.Veri giriş programı ekran görüntüsü
(Data entry program screenshot)

KOD	AD	OLC1
1111102270	Makarna (Restoranda) (Porsiyon - 4370)	4370
111601010	Makarna (Çubuk, Fiyonk, Erişte vb.) (Sade) (Kilogram - 1500)	1500
111601020	Makarna (Çubuk, Fiyonk, Erişte vb.) (Kepekli, Sebzeli, Sütlü, Yumurtalı vb.) (Kilogram - 1500)	1500
111601030	Makarna Spagetti (Sade) (Kilogram - 1500)	1500
111601040	Makarna Spagetti (Kepekli, Sebzeli, Sütlü, Yumurtalı vb.) (Kilogram - 1500)	1500
111601050	Makarna Kuskus (Kilogram - 1500)	1500
111601060	Makarna Lazanya (Kilogram - 1500)	1500
111601070	Makarna Kesme (Ev Makarnası-Erişte) (Kilogram - 1500)	1500
111601080	Makarna Fırın (Kilogram - 1500)	1500
111601090	Makarna Diyet (Kilogram - 1500)	1500

Şekil 2.Veri giriş programı ekran görüntüsü
(Data entry program screenshot)

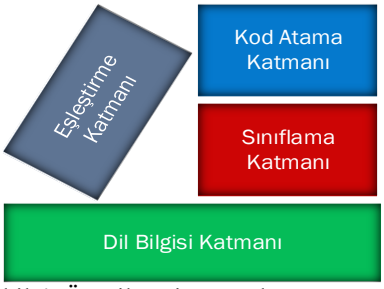
Harcamanın adı, cinsi ve tanımı	Markası	Kodu	Miktarı	Ölçü birimi	Toplam değeri	Piyasa birim...	
ekmek normal		111301010	0,6	4470	3	5,0	marki
bira teneke kutu alkollü 50 cl	efes	213101020	2	2600	18	9,0	marki
lokantada yenen tabldot yem		1111118010	1	4520	8	8,0	lokan
yaş pasta		111401010	1	2600	50	50,0	pasta
apartman ortak harcamaları		444101010	1	4560	50	50,0	apart
makarna		111601030					

Şekil 3.Veri giriş programı ekran görüntüsü
(Data entry program screenshot)

Ancak, "makarna" tanımı ile verilebilecek alternatif 10 kod olmasına rağmen bu kodların içerisinden anketör tarafından bir tanesi seçilmiştir. Seçim aşamasında anketör gerçekten bu harcamanın, verdiği koda uygun olduğunu bilebilir ama bu seçimini garanti altına alacak yeteri kadar ayrıntılı tanımı harcamanın adı, cinsi ve tanımı sütununa kaydetmesi gerekmektedir. Bu haliyle, daha sonra bu kaydı kontrol veya analiz eden kişinin bu kodun anketör tarafından doğru olarak verildiğini anlamasına ve bilmesine imkan bulunmamaktadır. Anketörün, makarna tanımıyla hanenin harcamasına uygun doğru kodu vermiş olma olasılığı 1/10'dur.

4.1. Önerilen Sistem (Proposed System)

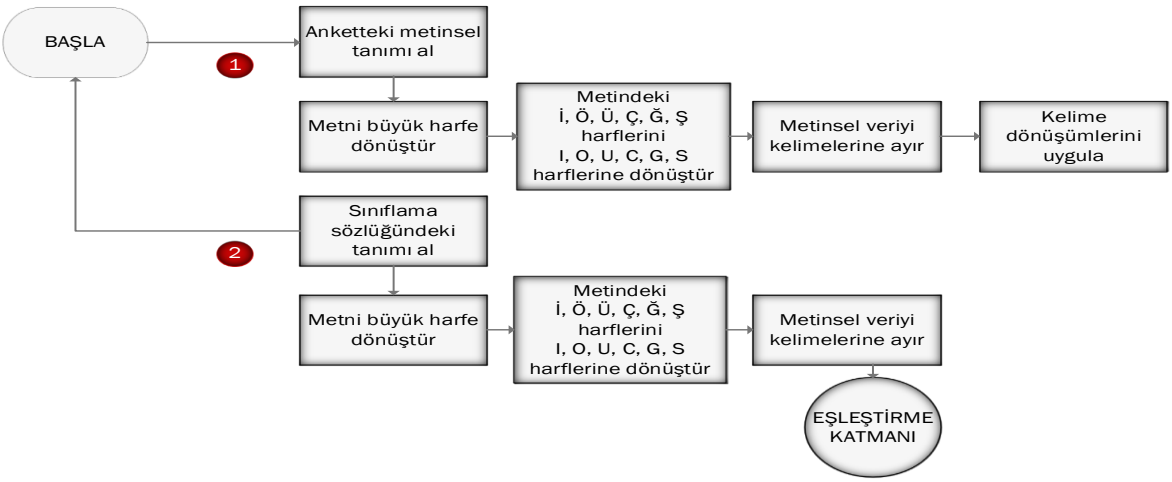
Sistem, SAS (Statistical Analysis Software) 9.3 W32_7PRO platformunda istatistiksel analiz için kullanılan bir bilgisayar programlama dili olan SAS dilinde geliştirilmiştir. Bu çalışmada geliştirilen kod atama sistemi için Şekil 4'te görüleceği üzere dört farklı katman bulunmaktadır.



Şekil 4. Önerilen sistemin katman yapısı
(Layer structure of proposed system)

4.1.1. Dil Bilgisi Katmanı (Grammar Layer)

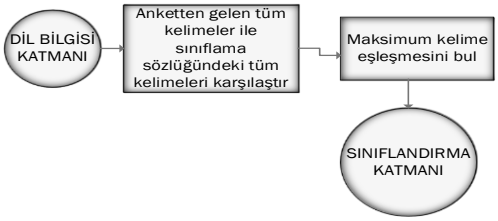
Bu katmanların birincisi, dil bilgisi katmanıdır. Şekil 5'te yer alan katmanda anketör tarafından yazılmış harcama tanımı büyük harfe dönüştürülür ve Türkçe karakterlerden arındırılır. Aynı işlemler sınıflama sözlüğündeki tanımlara da yapılarak sözlükte yer alan ve anketör tarafından yazılmış harcama tanımları kelimelerine ayrılır. Bu kelimelere, yazım yanlışlarını gidermek amacıyla daha önce belirlenmiş olan kelime dönüşümleri uygulanır.



Şekil 5. Dil bilgisi katmanı
(Grammar layer)

4.1.2. Eşleştirme Katmanı (Matching Layer)

Katmanlardan ikincisi, Şekil 6'da verilen eşleştirme katmanıdır. Bu katmanda kelimelerine ayrılmış harcama tanımı ile kelimelerine ayrılmış sınıflama sözlüğündeki tanımlar karşılaştırılır. Anketörün yazdığı harcama tanımının kelimelerinin sözlükteki hangi kod veya kodların tanımının kelimelerini maksimum olarak içerdiği bulunur.



Şekil 6. Eşleştirme katmanı
(Matching layer)

4.1.3. Sınıflama Katmanı (Classification Layer)

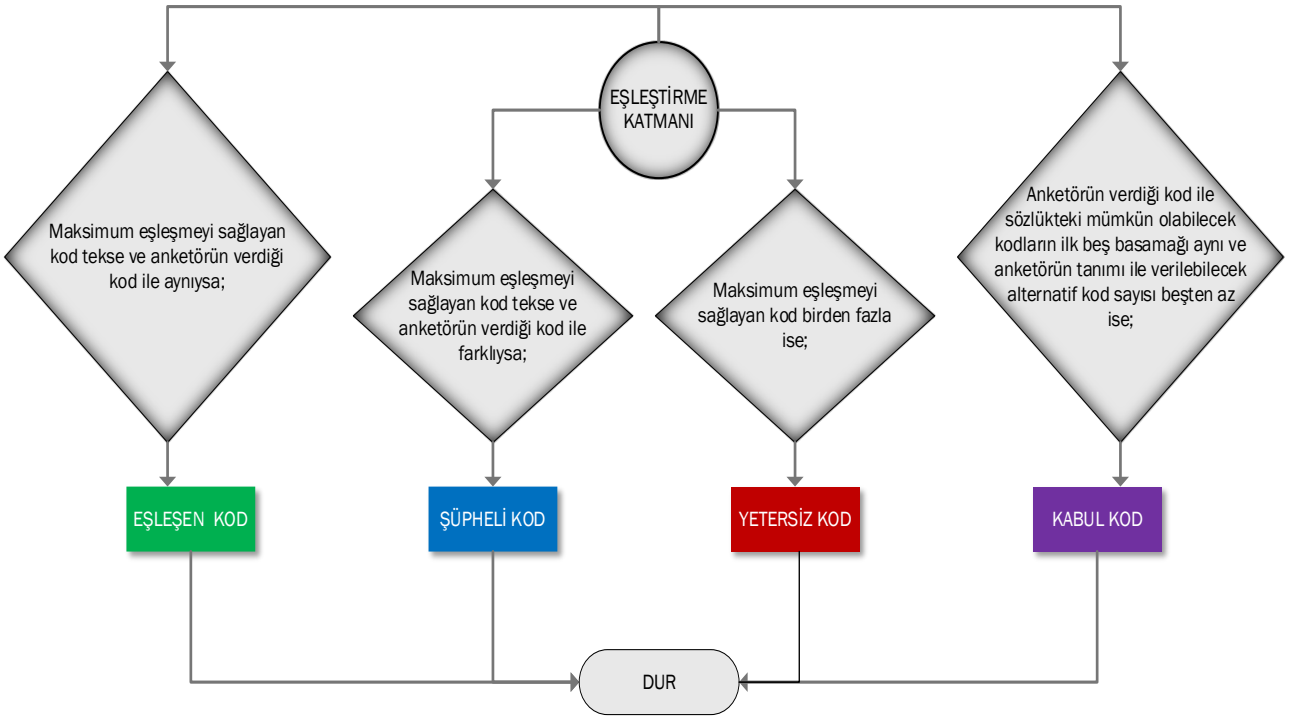
Katmanlardan üçüncüsü, sınıflama katmanıdır. Şekil 7'de yer alan bu katmanda maksimum kelime eşleşmeleri bulunan harcama tanımları ve sözlük tanımları sınıflanmaktadır. Burada dört sınıf bulunmaktadır:

1. **Eşleşen kod:** Anketörün yazdığı harcama tanımındaki kelimeleri en çok içeren sözlükteki kod tekse ve bu kod ile anketörün verdiği harcama kodu aynıysa.

2. **Şüpheli kod:** Anketörün yazdığı harcama tanımındaki kelimeleri en çok içeren standart sözlükteki kod tekse ve bu kod ile anketörün verdiği harcama kodu farklıysa.

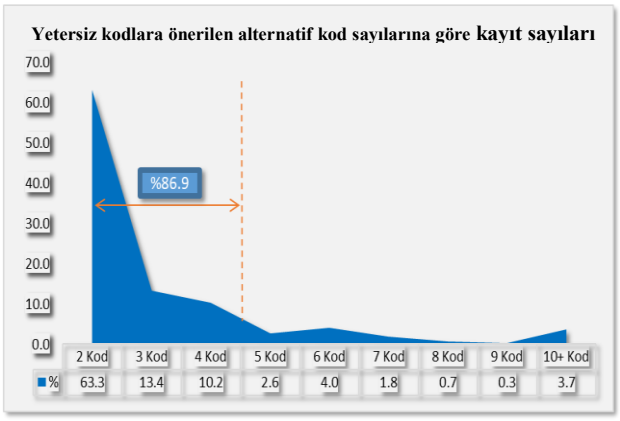
3. **Yetersiz kod:** Anketörün yazdığı harcama tanımındaki kelimeleri içeren standart sözlükteki kod birden fazlaysa.

4. **Kabul kod:** Anketörün verdiği kod sistem tarafından önerilen alternatif kodların arasında varsa, anketörün verdiği kod ile alternatif olarak önerilen kodların ilk beş basamağı aynı ve anketörün tanımı ile verilebilecek alternatif kod sayısı beşten azsa.



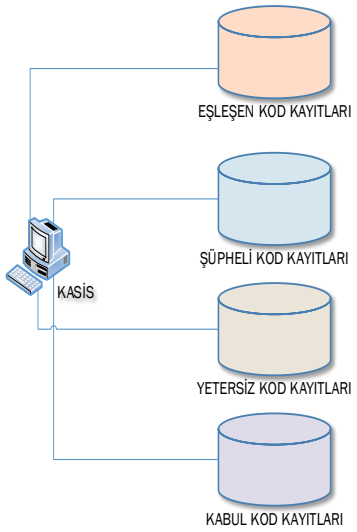
Şekil 7. Sınıflandırma katmanı
(Classification layer)

Kabul kod olarak sınıflaması yapılacak kayıtlar için belirlenmiş olan alternatif kod sayısının beşten az olma kriterine, veri setinde yetersiz kod olarak sınıflanmış harcama kayıtlarına sistemin önerdiği alternatif kod sayılarının Şekil 8'de verilen dağılıma göre karar verilmiştir. Program tarafından önerilen alternatif kod sayısının beşten az olma ve ilk beş basamağı aynı olma kriterlerini yetersiz kayıtların %86,9'u sağlamaktadır. Eurostat'a beş basamaklı tüketim harcamaları kodları ayrıntısında bilgiler gönderildiği için kodlarda bu ayrıntıda doğruluğun sağlanması bir kriter olarak dikkate alınmıştır.



Şekil 8. Yetersiz kodlara önerilen alternatif kod sayılarına göre kayıt sayıları
(The number of records by the number of alternative codes proposed for insufficient codes)

Sistem tarafından sınıflaması yapılan kayıtlar Şekil 9'da verilen veri tabanlarına ilave edilmektedir. Bu veri tabanlarından daha önce sınıflaması yapılan kayıtlara istenildiği zaman ulaşılabilmektedir. Ayrıca, veri tabanında daha önce mevcut olan kayıt tekrar geldiğinde sistemden geçmeden direkt olarak kaydın sınıflaması otomatik olarak yapılabilmektedir.



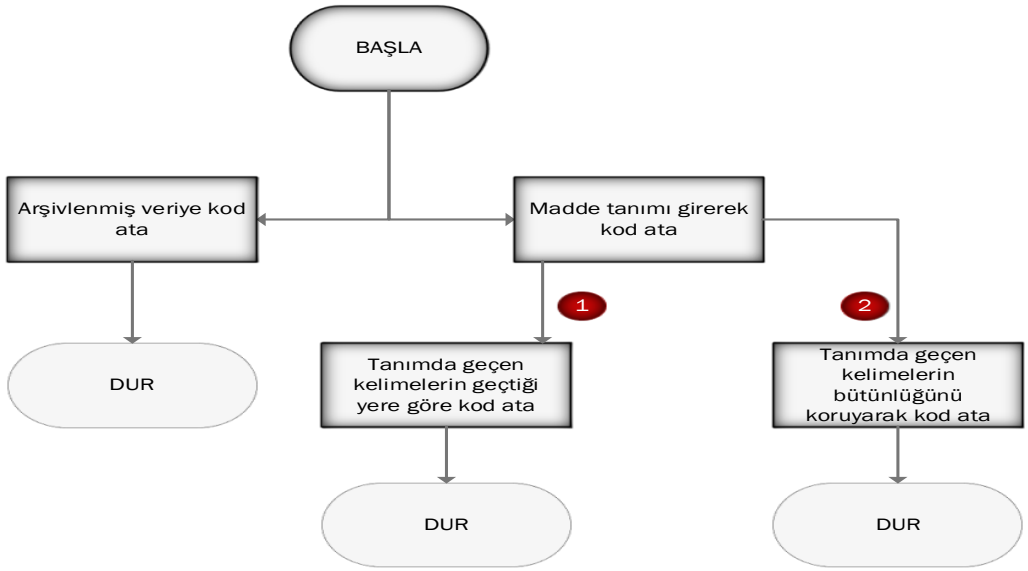
Şekil 9. Veri tabanı yapısı
(Database structure)

4.1.4. Kod Atama Katmanı (Code Assignment Layer)

Katmanlardan sonuncusu, kod atama katmanıdır. Şekil 10'da verilen bu katmanda daha önceki katmanlarda sınıflaması yetersiz olarak yapılmış kayıtlara bulanık eşleştirme (fuzzy matching) teknikleri ile kod ataması yapılabilmektedir. Ancak makale kapsamında bu konuya değinilmemiştir. Ayrıca madde tanımı manuel olarak girilebilen beş kayda harcama kod ataması da yapılabilmektedir. Burada iki seçenek bulunmaktadır.

1. **Tanımda geçen kelimelerin bütünlüğünü korumadan kod ata:** Bu atama yönteminde aranan kelime hangi sözlük tanımında bulunursa eşleştirme sağlanmaktadır. Örneğin "elma" harcama tanımının arandığı varsayarsak, "elmas" sözlük tanımının eşleşmesi gibi.

2. **Tanımda geçen kelimelerin bütünlüğünü koruyarak kod ata:** Bu atama yönteminde kelime bütünlüğü korunarak sözlük tanımında bulunursa eşleştirme sağlanmaktadır. Örneğin top harcama tanımının arandığı varsayarsak, "laptop" sözlük tanımının değil sadece "top" sözlük tanımının eşleşmesi gibi.



Şekil 10. Kod atama katmanı
(Code assignment layer)

5. BULGULAR (RESULTS)

Bu çalışmada, 2017 yılında TÜİK tarafından uygulanan HBA mikro veri setindeki daha önce anketör tarafından kodlaması yapılmış tüketim harcamaları kodlarının kontrolü geliştirilen sistem ile yapılmıştır. 2017 yılında Bölüm 4'te anlatılan tablolara göre kontrol edilen kayıt sayılarına Tablo 3'te yer verilmiştir.

Tablo 3. 2017 yılı verisindeki tablolara göre kayıt sayıları
(Number of records according to tables in 2017 data)

TABLO	KAYIT SAYISI	KAYIT SAYISI %
COICOP_AYNI	125.484	5,6
COICOP_DAYANIKLI	40.227	1,8
COICOP_SATINALIS	1.834.868	82,5
COICOP_STOK	224.593	10,1
Toplam	**2.225.172**	**100,0**

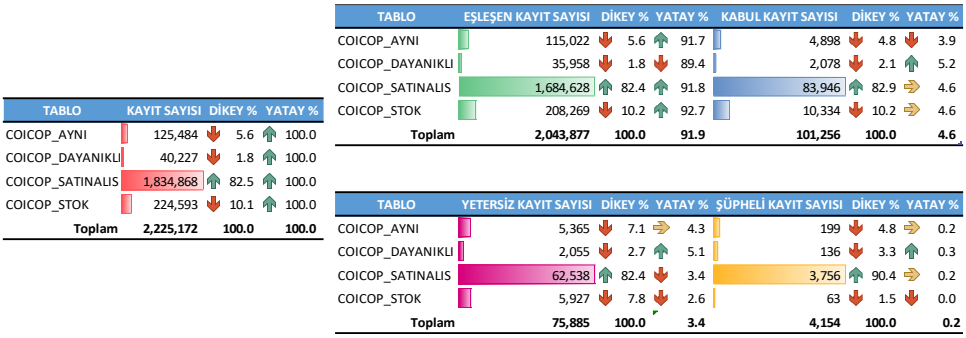
Sistem tarafından incelenen kodların tablolara göre dağılımına Şekil 11'de yer verilmiştir. Bu harcama kayıtlarına ait kodların %91,9'unda anketörün verdiği kod ile sistemin verdiği kodun aynı olduğu sonucuna ulaşılmıştır. Stok tablosu kodlarının %92,7'lik kısmında anketörün verdiği kod ile sistemin verdiği kod aynı iken dayanıklı tablosu kodlarının ise %89,4'lük kısmı bu sınıfta yer alabilmiştir.

Bölüm 4.1.3'te belirlenen kriterlere göre kodların %4,6'sı kabul kod olarak sınıflandırılmıştır. Bu sınıfa en büyük katkıyı satınalış tablosu kodlarının %4,6'lık kısmı yaparken en düşük katkıyı dayanıklı tablosu kodlarının %5,2'lik kısmı yapmıştır.

Kodların %3,6'lık bölümünün ise yeniden incelenmesi ve irdelenmesine ihtiyaç bulunmaktadır. Bu kodların %3,4'ü yetersiz olarak, %0,2'si şüpheli olarak sınıflandırılmıştır. Sistemin şüpheli olarak sınıflandırdığı yani anketörün verdiği kodun hatalı olduğu kayıtların mutlaka düzeltilmesi ve sistemin yetersiz olarak sınıflandırdığı yani sistemin yazılan tanıma göre birden fazla alternatif kodu bulduğu kayıtlarda tanımların ise biraz daha ayrıntılı olarak yazılması önerilmektedir. Dayanıklı tablosu kodlarının %5,1'i yetersiz ve %0,3'ü şüpheli kod olarak sınıflanarak diğer tablolardan ön plana çıkmaktadır.

Görüldüğü üzere, genel olarak verilen kodların sadece %3,6'lık kısmında sorun bulunmaktadır. Dolayısıyla bu kodlar kullanılarak yapılacak bir araştırma veya yayınlanacak sonuç bizi hatalı yerlere götürebilecektir. Bu

durumu engellemek adına iki tür çözüm önerisi sunulabilir. Bunlardan birincisi, sorunlu kodları dışarıda bırakarak çalışmalarımızı kodların geri kalan kısmıyla yapmaktır. Bu çözüm önerisi, çalışma sonuçlarının doğruluğunu bize garanti edecektir. İkinci çözüm önerisi, sistem tarafından şüpheli olarak sınıflandırılmış kodların sistem tarafından önerilmiş hallerini kullanmak ve yetersiz olarak sınıflandırılmış kodlarda ise sistem tarafından önerilen kodlar arasından bir seçim yapmaktır.



TABLO	KAYIT SAYISI	DİKEY %	YATAY %
COICOP_AYNI	125,484	5.6	100.0
COICOP_DAYANIKLI	40,227	1.8	100.0
COICOP_SATINALIS	1,834,868	82.5	100.0
COICOP_STOK	224,593	10.1	100.0
Toplam	2,225,172	100.0	100.0

TABLO	EŞLEŞEN KAYIT SAYISI	DİKEY %	YATAY %	KABUL KAYIT SAYISI	DİKEY %	YATAY %
COICOP_AYNI	115,022	5.6	91.7	4,898	4.8	3.9
COICOP_DAYANIKLI	35,958	1.8	89.4	2,078	2.1	5.2
COICOP_SATINALIS	1,684,628	82.4	91.8	83,946	82.9	4.6
COICOP_STOK	208,269	10.2	92.7	10,334	10.2	4.6
Toplam	2,043,877	100.0	91.9	101,256	100.0	4.6

TABLO	YETERSİZ KAYIT SAYISI	DİKEY %	YATAY %	ŞÜPHELİ KAYIT SAYISI	DİKEY %	YATAY %
COICOP_AYNI	5,365	7.1	4.3	199	4.8	0.2
COICOP_DAYANIKLI	2,055	2.7	5.1	136	3.3	0.3
COICOP_SATINALIS	62,538	82.4	3.4	3,756	90.4	0.2
COICOP_STOK	5,927	7.8	2.6	63	1.5	0.0
Toplam	75,885	100.0	3.4	4,154	100.0	0.2

Şekil 11. Tablolara göre 2017 yılı sonuçlarının dağılımı
(Distribution of 2017 results by tables)

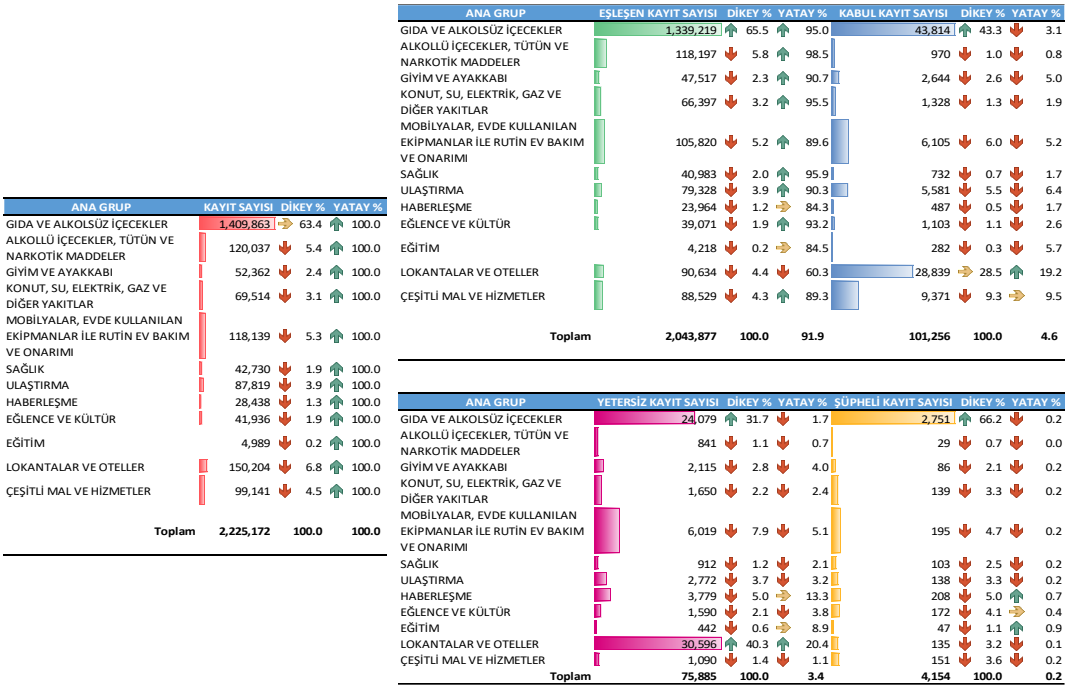
Sistem tarafından incelenen kodların Şekil 12'de verilen ana grup dağılımlarına göre bakıldığı zaman, kodların %63,4'ünün gıda ve alkolsüz içecekler grubunda olduğu görülmektedir. En düşük pay %0,2 ile eğitim grubu kodlarına aittir.

Lokanta ve oteller ana grubu kayıtlarının %60,3'lük kısmı eşleşen kod olarak sınıflandırılmasına rağmen alkollü içecekler, tütün ve narkotik maddeler ana grubu kayıtlarının %98,5'lik kısmı eşleşen kod olarak sınıflandırılmıştır. Yani anketörün verdiği kod ile sistemin önerdiği kod birebir tutarlıdır.

Bölüm 4.1.3'te belirlenen kriterlere göre, lokanta ve oteller ana grubu kayıtlarının %19,2'lik kısmı kabul kod sınıfında yer almasına rağmen alkollü içecekler, tütün ve narkotik maddeler ana grubu kayıtlarının %0,8'lik kısmı bu sınıfta yer almıştır.

Alkollü içecekler, tütün ve narkotik maddeler ana grubu kayıtlarının %0,7'lik kısmı yetersiz kod olarak sınıflandırılmasına rağmen lokantalar ve oteller ana grubu kayıtlarının %20,4'lik kısmı bu sınıfta yer almıştır. Bu sınıfta yer alan kodların tanımının anketör tarafından biraz daha ayrıntılı yazılması gerekmektedir.

Alkollü içecekler, tütün ve narkotik maddeler ana grubu kayıtlarından sadece 29 adedi kodu şüpheli olarak sınıflandırılmıştır. Yani anketörün verdiği kod ile sistemin önerdiği kod farklıdır. Bu kod ve tanımların tekrar incelenmesine ihtiyaç bulunmaktadır. Aynı sınıfta eğitim ana grubu kayıtlarının %0,9'luk kısmı yer almıştır.



ANA GRUP	KAYIT SAYISI	DİKEY %	YATAY %
GIDA VE ALKOLSÜZ İÇECEKLER	1,409,863	63.4	100.0
ALKOLLÜ İÇECEKLER, TÜTÜN VE NARKOTİK MADDELER	120,037	5.4	100.0
GİYİM VE AYAKKABI	52,362	2.4	100.0
KONUT, SU, ELEKTRİK, GAZ VE DİĞER YAKITLAR	69,514	3.1	100.0
MOBİLYALAR, EVDE KULLANILAN EKİPMANLAR İLE RUTİN EV BAKIM VE ONARIMI	118,139	5.3	100.0
SAĞLIK	42,730	1.9	100.0
ULAŞTIRMA	87,819	3.9	100.0
HABERLEŞME	28,438	1.3	100.0
EĞLENCE VE KÜLTÜR	41,936	1.9	100.0
EĞİTİM	4,989	0.2	100.0
LOKANTALAR VE OTELLER	150,204	6.8	100.0
ÇEŞİTLİ MAL VE HİZMETLER	99,141	4.5	100.0
Toplam	2,225,172	100.0	100.0

ANA GRUP	EŞLEŞEN KAYIT SAYISI	DİKEY %	YATAY %	KABUL KAYIT SAYISI	DİKEY %	YATAY %
GIDA VE ALKOLSÜZ İÇECEKLER	1,339,219	65.5	95.0	43,814	43.3	3.1
ALKOLLÜ İÇECEKLER, TÜTÜN VE NARKOTİK MADDELER	118,197	5.8	98.5	970	1.0	0.8
GİYİM VE AYAKKABI	47,517	2.3	90.7	2,644	2.6	5.0
KONUT, SU, ELEKTRİK, GAZ VE DİĞER YAKITLAR	66,397	3.2	95.5	1,328	1.3	1.9
MOBİLYALAR, EVDE KULLANILAN EKİPMANLAR İLE RUTİN EV BAKIM VE ONARIMI	105,820	5.2	89.6	6,105	6.0	5.2
SAĞLIK	40,983	2.0	95.9	732	0.7	1.7
ULAŞTIRMA	79,328	3.9	90.3	5,581	5.5	6.4
HABERLEŞME	23,964	1.2	84.3	487	0.5	1.7
EĞLENCE VE KÜLTÜR	39,071	1.9	93.2	1,103	1.1	2.6
EĞİTİM	4,218	0.2	84.5	282	0.3	5.7
LOKANTALAR VE OTELLER	90,634	4.4	60.3	28,839	28.5	19.2
ÇEŞİTLİ MAL VE HİZMETLER	88,529	4.3	89.3	9,371	9.3	9.5
Toplam	2,043,877	100.0	91.9	101,256	100.0	4.6

ANA GRUP	YETERSİZ KAYIT SAYISI	DİKEY %	YATAY %	ŞÜPHELİ KAYIT SAYISI	DİKEY %	YATAY %
GIDA VE ALKOLSÜZ İÇECEKLER	24,079	31.7	1.7	2,751	66.2	0.2
ALKOLLÜ İÇECEKLER, TÜTÜN VE NARKOTİK MADDELER	841	1.1	0.7	29	0.7	0.0
GİYİM VE AYAKKABI	2,115	2.8	4.0	86	2.1	0.2
KONUT, SU, ELEKTRİK, GAZ VE DİĞER YAKITLAR	1,650	2.2	2.4	139	3.3	0.2
MOBİLYALAR, EVDE KULLANILAN EKİPMANLAR İLE RUTİN EV BAKIM VE ONARIMI	6,019	7.9	5.1	195	4.7	0.2
SAĞLIK	912	1.2	2.1	103	2.5	0.2
ULAŞTIRMA	2,772	3.7	3.2	138	3.3	0.2
HABERLEŞME	3,779	5.0	13.3	208	5.0	0.7
EĞLENCE VE KÜLTÜR	1,590	2.1	3.8	172	4.1	0.4
EĞİTİM	442	0.6	8.9	47	1.1	0.9
LOKANTALAR VE OTELLER	30,596	40.3	20.4	135	3.2	0.1
ÇEŞİTLİ MAL VE HİZMETLER	1,090	1.4	1.1	151	3.6	0.2
Toplam	75,885	100.0	3.4	4,154	100.0	0.2

Şekil 12. Ana gruplara göre 2017 yılı sonuçlarının dağılımı
(Distribution of 2017 results by main groups)

Şekil 13'te anket yapılan aylara göre sonuçların dağılımı verilmiştir. Bu sonuçlara göre, Aralık ayı kayıtlarının %93,2'si eşleşen olarak sınıflandırılmış olmasına rağmen Ocak ayı kayıtlarının %91'i bu sınıfta yer almıştır. Yani, bölüm 4.1.3'te belirlenen kriterlere göre, anketörün verdiği kod ile sistemin önerdiği birebir aynıdır.

Bölüm 4.1.3'te belirlenen kriterlere göre, Ocak ayı kayıtlarının %5'i kabul kod olarak sınıflandırılmış olmasına rağmen Aralık ayı kayıtlarının %4,1'i bu sınıfta yer almıştır.

Aralık ayı kayıtlarının %2,5'i yetersiz kayıt olarak sınıflandırılırken Temmuz ve Ağustos ayı kayıtlarının %3,9'u bu sınıfta yer almıştır. Bir başka deyişle, anketörün yazdığı tanım verdiği kodu verebilecek kadar detaylı yazılmalıdır.

REFERANS AY	KAYIT SAYISI	DİKEY %	YATAY %
1	185,345	8.3	100.0
2	174,903	7.9	100.0
3	189,758	8.5	100.0
4	191,189	8.6	100.0
5	192,786	8.7	100.0
6	173,086	7.8	100.0
7	173,410	7.8	100.0
8	177,130	8.0	100.0
9	180,440	8.1	100.0
10	199,434	9.0	100.0
11	198,142	8.9	100.0
12	189,549	8.5	100.0
Toplam	2,225,172	100.0	100.0

REFERANS AY	EŞLEŞEN KAYIT SAYISI	DİKEY %	YATAY %	KABUL KAYIT SAYISI	DİKEY %	YATAY %
1	168,718	8.3	91.0	9,299	9.2	5.0
2	159,877	7.8	91.4	8,116	8.0	4.6
3	173,717	8.5	91.5	9,359	9.2	4.9
4	175,417	8.6	91.8	8,614	8.5	4.5
5	176,914	8.7	91.8	9,063	9.0	4.7
6	159,602	7.8	92.2	7,645	7.6	4.4
7	158,593	7.8	91.5	7,671	7.6	4.4
8	161,918	7.9	91.4	8,018	7.9	4.5
9	166,185	8.1	92.1	7,908	7.8	4.4
10	183,597	9.0	92.1	8,842	8.7	4.4
11	182,631	8.9	92.2	8,944	8.8	4.5
12	176,708	8.6	93.2	7,777	7.7	4.1
Toplam	2,043,877	100.0	91.9	101,256	100.0	4.6

REFERANS AY	YETERSİZ KAYIT SAYISI	DİKEY %	YATAY %	ŞÜPHELİ KAYIT SAYISI	DİKEY %	YATAY %
1	6,995	9.2	3.8	333	8.0	0.2
2	6,613	8.7	3.8	297	7.1	0.2
3	6,345	8.4	3.3	337	8.1	0.2
4	6,801	9.0	3.6	357	8.6	0.2
5	6,460	8.5	3.4	349	8.4	0.2
6	5,523	7.3	3.2	316	7.6	0.2
7	6,696	8.8	3.9	450	10.8	0.3
8	6,836	9.0	3.9	358	8.6	0.2
9	6,019	7.9	3.3	328	7.9	0.2
10	6,668	8.8	3.3	327	7.9	0.2
11	6,181	8.1	3.1	386	9.3	0.2
12	4,748	6.3	2.5	316	7.6	0.2
Toplam	75,885	100.0	3.4	4,154	100.0	0.2

Şekil 13. Aylara göre 2017 yılı sonuçlarının dağılımı
(Distribution of 2017 results by months)

İstatistiki Bölge Birimleri Sınıflaması (İBBS) Düzey 2 bölgelerine ait sonuçlar Şekil 14'te verilmiştir. Bu sonuçlara göre, Diyarbakır ve Şanlıurfa illerini kapsayan Diyarbakır bölgesi kayıtlarının %82,2'si eşleşen olarak sınıflandırılmış olmasına rağmen Ankara ilini kapsayan Ankara ve Kastamonu, Çankırı ve Sinop illerini kapsayan Kastamonu bölgesi kayıtlarının %96,8'i bu sınıfta yer almıştır. Yani anketörün verdiği kod ile sistemin önerdiği kod ile birebir tutarlıdır. Kastamonu, Çankırı ve Sinop illerini kapsayan Kastamonu bölgesi kayıtlarının %2,1'i bölüm 4.1.3'te belirlenen kriterlere göre kabul kod olarak sınıflandırılmış olmasına rağmen Diyarbakır ve Şanlıurfa illerini kapsayan Diyarbakır bölgesi kayıtlarının Diyarbakır bölgesi kayıtlarının %10,9'u bu sınıfta yer almıştır. Erzurum, Erzincan ve Bayburt illerini kapsayan Erzurum bölgesi kayıtlarının %9,6'sı ve Ankara ilini kapsayan Ankara %0,8'i tanımı yetersiz olarak sınıflandırılmıştır. Kars, Ağrı, Iğdır ve Ardahan illerini kapsayan Kars bölgesi kayıtlarının %1,1'i şüpheli olarak sınıflandırılırken, Gaziantep, Adıyaman ve Kilis illerini kapsayan Gaziantep bölgesi kayıtlarından sadece 10 tanesi bu sınıfta yer almıştır. Bu kodların düzeltilmesi önerilmektedir.

BÖLGE KOD	KAYIT SAYISI	DİKEY %	YATAY %
ADANA	131,306	5.9	100.0
ANKARA	204,104	9.2	100.0
ANTALYA	76,431	3.4	100.0
BALIKESİR	62,783	2.8	100.0
BURSA	134,820	6.1	100.0
DENİZLİ	111,984	5.0	100.0
DİYARBAKIR	23,352	1.0	100.0
EDİRNE	71,241	3.2	100.0
ERZURUM	43,631	2.0	100.0
GAZİANTEP	64,990	2.9	100.0
HATAY	78,741	3.5	100.0
İSTANBUL	335,054	15.1	100.0
İZMİR	151,225	6.8	100.0
KARS	30,427	1.4	100.0
KASTAMONU	31,211	1.4	100.0
KAYSERİ	69,710	3.1	100.0
KOCAELİ	98,584	4.4	100.0
KONYA	78,423	3.5	100.0
MALATYA	50,498	2.3	100.0
MANİSA	91,459	4.1	100.0
NEVŞEHİR	35,856	1.6	100.0
SAMSUN	69,584	3.1	100.0
SİİRT	33,317	1.5	100.0
TRABZON	60,597	2.7	100.0
VAN	43,921	2.0	100.0
ZONGULDAK	41,923	1.9	100.0
Toplam	**2,225,172**	**100.0**	**100.0**

BÖLGE KOD	EŞLEŞEN KAYIT SAYISI	DİKEY %	YATAY %	KABUL KAYIT SAYISI	DİKEY %	YATAY %
ADANA	123,218	6.0	93.8	5,479	5.4	4.2
ANKARA	197,658	9.7	96.8	4,681	4.6	2.3
ANTALYA	69,420	3.4	90.8	4,278	4.2	5.6
BALIKESİR	54,528	2.7	86.9	5,597	5.5	8.9
BURSA	123,268	6.0	91.4	6,811	6.7	5.1
DENİZLİ	96,689	4.7	86.3	6,519	6.4	5.8
DİYARBAKIR	19,184	0.9	82.2	2,551	2.5	10.9
EDİRNE	66,187	3.2	92.9	2,978	2.9	4.2
ERZURUM	36,235	1.8	83.0	3,046	3.0	7.0
GAZİANTEP	58,607	2.9	90.2	4,272	4.2	6.6
HATAY	75,389	3.7	95.7	2,631	2.6	3.3
İSTANBUL	313,502	15.3	93.6	11,218	11.1	3.3
İZMİR	138,019	6.8	91.3	6,884	6.8	4.6
KARS	26,429	1.3	86.9	2,031	2.0	6.7
KASTAMONU	30,210	1.5	96.8	665	0.7	2.1
KAYSERİ	65,699	3.2	94.2	3,031	3.0	4.3
KOCAELİ	88,815	4.3	90.1	4,786	4.7	4.9
KONYA	65,636	3.2	83.7	5,293	5.2	6.7
MALATYA	41,867	2.0	82.9	5,473	5.4	10.8
MANİSA	83,521	4.1	91.3	3,874	3.8	4.2
NEVŞEHİR	34,330	1.7	95.7	906	0.9	2.5
SAMSUN	66,945	3.3	96.2	1,547	1.5	2.2
SİİRT	30,091	1.5	90.3	1,614	1.6	4.8
TRABZON	57,893	2.8	95.5	1,943	1.9	3.2
VAN	41,481	2.0	94.4	1,425	1.4	3.2
ZONGULDAK	39,056	1.9	93.2	1,723	1.7	4.1
Toplam	**2,043,877**	**100.0**	**91.9**	**101,256**	**100.0**	**4.6**

BÖLGE KOD	YETERSİZ KAYIT SAYISI	DİKEY %	YATAY %	ŞÜPHELİ KAYIT SAYISI	DİKEY %	YATAY %
ADANA	2,536	3.3	1.9	73	1.8	0.1
ANKARA	1,698	2.2	0.8	67	1.6	0.0
ANTALYA	2,595	3.4	3.4	138	3.3	0.2
BALIKESİR	2,615	3.4	4.2	43	1.0	0.1
BURSA	4,504	5.9	3.3	237	5.7	0.2
DENİZLİ	8,109	10.7	7.2	667	16.1	0.6
DİYARBAKIR	1,590	2.1	6.8	27	0.6	0.1
EDİRNE	2,043	2.7	2.9	33	0.8	0.0
ERZURUM	4,209	5.5	9.6	141	3.4	0.3
GAZİANTEP	2,101	2.8	3.2	10	0.2	0.0
HATAY	680	0.9	0.9	41	1.0	0.1
İSTANBUL	9,791	12.9	2.9	543	13.1	0.2
İZMİR	5,996	7.9	4.0	326	7.8	0.2
KARS	1,621	2.1	5.3	346	8.3	1.1
KASTAMONU	283	0.4	0.9	53	1.3	0.2
KAYSERİ	933	1.2	1.3	47	1.1	0.1
KOCAELİ	4,755	6.3	4.8	228	5.5	0.2
KONYA	7,344	9.7	9.4	150	3.6	0.2
MALATYA	2,923	3.9	5.8	235	5.7	0.5
MANİSA	3,924	5.2	4.3	140	3.4	0.2
NEVŞEHİR	610	0.8	1.7	10	0.2	0.0
SAMSUN	1,005	1.3	1.4	87	2.1	0.1
SİİRT	1,405	1.9	4.2	207	5.0	0.6
TRABZON	704	0.9	1.2	57	1.4	0.1
VAN	823	1.1	1.9	192	4.6	0.4
ZONGULDAK	1,088	1.4	2.6	56	1.3	0.1
Toplam	**75,885**	**100.0**	**3.4**	**4,154**	**100.0**	**0.2**

Şekil 14. İBBS Düzey 2 bölgelerine göre 2017 yılı sonuçlarının dağılımı
(Distribution of 2017 results according to NUTS 2)

Otomatik kodlama programlarının performansını karşılaştırmak için genellikle kullanılan kriterler; verimlilik, güvenilirlik ve hız olmaktadır. Ancak, bu kriterler mutlak değildir [10].

Çalışma kapsamında daha önce anketör tarafından kodlaması yapılarak sonuçları kamuoyu ile paylaşılmış olan veri üzerinde sistemin etkinliği test edilmiştir. Kod Atama Sistemi (KASİS) ile kayıtların %96,5'ine anketör ile aynı kod ataması yapılarak bu kayıtlarda yapılan kodlamanın doğru olduğu sonucuna varılmıştır.

Roessingh ve Bethlehem [19], aile harcama anketinde üç farklı yöntemle otomatik kod ataması gerçekleştirmişlerdir ve %94, %85, %78 oranlarında doğru kodlama gerçekleştirmişlerdir.

Yeni Zelanda İstatistik Ofisi, Census 2013 verilerinde meslek ve okul sonrası yeterlilik değişkenlerini kodlamak için SVM algoritmasını kullanmışlardır. Her iki değişken için test verilerinde %50 doğruluk oranına ulaşmışlardır [26].

Tourigny ve Moloney [14], 1991 Kanada Nüfus Sayımı'nda yer alan yedi farklı değişken için yapılan otomatik kodlama sonucunda %92'lik bir doğruluk oranına ulaşmışlardır.

Haslinger [31], Avusturya Nüfus Sayımı'nda yer alan çalışılan yer değişkenine %96, eğitim değişkenine %92 ve iktisadi faaliyet değişkenine %50 oranında otomatik kodlama ile kod ataması yapabilmiştir.

6. SONUÇLAR (CONCLUSION)

Sınıflamalar, verilerin belirli standartlara uygun olarak toplanmasını ve analiz edilmesini sağlayan araçlardır [4]. Doğru sınıflama için ankette derlenen metinsel cevapları, önceden tanımlanmış kod yapısına doğru aktarmak gerekmektedir. Ankete cevap veren kişiden alınan metinsel ifadelerin kodlara dönüştürülmesi esnasında insan faktörü devreye girdiği için alınan metinsel ifade doğru olsa bile atanan sınıflama kodu hatalı olabilmektedir. Bu bakımdan, farklı toplumlarda farklı değer yargılarına ve algılarına sahip insanların yaşadığı da göz önünde bulundurularak ve insandan uzaklaşan ve insan etkisinin olmadığı analiz süreçleri ile sistematik olarak çalışan bir kod atama sisteminin geliştirilmesi önem arz etmektedir. Bu yüzden sadece kod verenin inisiyatifi ile sonuçlar değerlendirilmemeli, verilen kodların sınıflama sözlüğüne uygun olarak verilip verilmediğinin kontrol edilmesi ve sonuçlarının analiz edilmesi gerekmektedir. Bu kontrolün manuel olarak yapılması kontrol edilecek kayıt sayısının artması ile birlikte zamanlılık, maliyet ve kalite düşünüldüğünde çok verimli olmayacaktır. Ayrıca oluşturulan sisteminin standart sınıflama sözlüğü olan gelecekte oluşacak ve şu an mevcut diğer sınıflamalar da uygulanabilir olması son derece önemlidir. Bu bakımdan geliştirilen bu otomatik kod atama sistemi verimlilik, doğruluk ve hız bakımından yapılacak çalışmalara katkı sağlayacaktır.

Otomatik kodlama uygulamalarında genellikle denetimli makine öğrenmesi yöntemleri kullanılmaktadır. Denetimli makine öğrenmesinin başarısı büyük ölçüde bağımsız değişkenlerin tahmin gücüne ve eğitim veri setinin boyutuna bağlıdır. Yani, eğitim veri setinin hacmi ne kadar büyük ise modelin öğrenmesi o denli iyi olmaktadır [15]. Bu tip uygulamalarda modelin iyi öğrenebilmesi için kullanılacak eğitim veri setinin de doğruluğunun teyit edilmesi gerekmektedir. Aksi halde, yanlış öğrenen model yanlış sonuçlar üretecektir.

Geliştirilen bu sistemin diğer sistemler üzerindeki en önemli üstünlüğü eğitim veri setine ihtiyaç duymamasıdır. Sistem, sıfır noktasından itibaren kayıt sayısı arttıkça öğrenmesini de artırmaktadır. Bu sistemin diğer sistemlerden ayrılan üç yönden üstünlüğü bulunmaktadır. İlki, bu sistem makine öğrenmesinde kullanılacak eğitim veri setinin doğruluğunun teyit edilmesinde ve temizlenmesinde kullanılabilir. İkincisi, bu sistem diğer sistemlerin yaptığı gibi otomatik kod atama sistemidir. Sonuncusu ise, bu sistemin biriktirdiği eşleşen kayıtlar bir başka denetimli makine öğrenme uygulamasında eğitim veri seti olarak kullanılabilir. Bu çalışma, hangi anket yöntemi kullanılırsa kullanılsın anketörler veya kodlayıcılar tarafından kodlanmış kayıtların kalitesini artıracaktır. Sistem, kodlama tutarlılığını ve hassasiyetini artırarak anket maliyetlerini ve anketörün kodlama esnasında oluşturduğu görüşme yükünü azaltacaktır.

Çalışma kapsamında, anketör tarafından kodlaması yapılmış 2017 HBA verisi üzerinde sistemin etkinliği test edilmiştir. KASİS, kayıtların %96,5'inde anketör ile aynı sonuca ulaşmıştır.

Otomatik kodlama sisteminin doğru kodu atama performansı, sınıflamanın ve verilerin karmaşıklığına bağlı olarak değişebilmektedir. Ancak genel bir perspektif sunabilmesi açısından, KASİS'in kod atama performansı benzer sistemler ile karşılaştırıldığında başarılı olduğu sonucuna ulaşılmıştır [14, 19, 26, 31]. Kontrol edilen kayıtların %3,5'luk kısmı kodların doğruluğu bakımından yeniden irdelenmelidir.

Bundan sonraki süreçte, ilk hedefimiz KASİS'in direkt olarak istatistik üretim aşamasında kullanılmasını sağlamak olacaktır. Bu sayede, anket sonuçları kamuoyu ile paylaşılmadan önce kodlamada yapılan hata ve eksiklikler giderilmiş, kodlama kalitesi artırılmış ve sonucunda maliyetlerde azalma sağlanmış olacaktır.

İkinci hedefimiz, yazılan harcama tanımlarına direkt olarak KASİS tarafından kod ataması yapılmasını sağlayarak kodlama konusunda anketörlerin üzerindeki iş yükünü azaltmak olacaktır.

KAYNAKLAR (REFERENCES)

[1] Internet: Kalite Güven Çerçevesi 2015, Türkiye İstatistik Kurumu, http://www.tuik.gov.tr/jsp/duyuru/upload/TUIK_Kalite_Guvence_Cercevesi.pdf, 10.04.2019.

[2] F. Akdeniz, "İstatistikte Yeni Eğilimler ve Yöntemler", *Journal of Statistical Research*, 10(3), 35-48, 2013.

[3] Internet: Sınıflama Sunucusu, Türkiye İstatistik Kurumu, https://biruni.tuik.gov.tr/DIESS/, 10.04.2019.

[4] Internet: NACE Rev.2 Altılı Ekonomik Faaliyet Sınıflaması, Türkiye Odalar ve Borsalar Birliği, http://gen.tobb.org.tr/ggnot/images/bilgi_notu/277_DUYURU-GD1.pdf, 14.04.2019.

[5] Internet: Sınıflama Sunucusu, https://biruni.tuik.gov.tr/DIESS/SiniflamaTurListeAction.do, 12.04.2019.

[6] Internet: Türk Dil Kurumu Büyük Türkçe Sözlük, http://www.tdk.gov.tr/index.php?option=com_bts&arama=kelime&guid=TDK.GTS.59e33dfc8b1280.30675020, 10.04.2019.

[7] Internet: Quality Guidelines for Official Statistics, https://unstats.un.org/unsd/dnss/docs-nqaf/Finland-g_2ed_en.pdf, 10.04.2019.

[8] Internet: Türkiye İstatistik Kanunu, http://tuik.gov.tr/jsp/duyuru/upload/TuikKanun.pdf, 12.04.2019.

[9] M. Schierholz, **Automating survey coding for occupation,** Yüksek Lisans Tezi, Ludwig Maximilians Universitat, Institut fur Statistik, 2014.

[10] W. Hacking, L. Willenborg, **Theme: Coding; Interpreting Short Descriptions Using a Classification**, Statistics Netherlands, The Hague/Heerlen, 2012.

[11] C. C. Aggarwal, C. A. Zhai, "Survey of Text Classification Algorithms", **Mining Text Data**, Springer, Boston, MA, 163-222, 2012.

[12] R. M. Groves, F. J. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer, R. Tourangeau, **Survey Methodology**, John Wiley & Sons, New Jersey, A.B.D., 2009.

[13] A. Esuli, F. Sebastiani, "Machines That Learn How To Code Open-Ended Survey Data", *International Journal of Market Research*, 52(6), 775-800, 2010.

[14] J. Y. Tourigny, J. Moloney, **Statistical Data Editing Volume No. 2 Methods And Techniques**, United Nations Statistical Commission and Economic Commission for Europe, New York and Geneva, 1997.

[15] P. Dalton, G. Keogh. "Automatic Coding of Occupations: The Irish Experience". **New Techniques and Technologies for Statistics II Proceedings of the Second Bonn Seminar**, Netherlands: IOS Press, 33-44, 1997

[16] F. G. Conrad, "Using Expert Systems to Model and Improve Survey Classification Processes", **Survey Measurement and Process Quality**, John Wiley & Sons, New York, 393–414, 1997.

[17] D. Bushnell, "An Evaluation of Computer-assisted Occupation Coding", **New Methods for Survey Research**, Southampton, 23–36, 1998.

[18] J. Fielding, N. Fielding, G. Hughes, "Opening Up Open-Ended Survey Data Using Qualitative Software", *Quality & Quantity*, 47(6), 3261–3276, 2013.

[19] M. Roessingh, J. Bethlehem, "Trigram coding in the family expenditure survey in statistics," Netherlands Central Bureau of Statistics, 1983.

[20] F. R. Clarke, S. Brooker, "Use of Machine Learning for Automated Survey Coding", **International Statistical Institute Proceedings of the 58th World Statistics Congress 2011**, Dublin Convention Centre, İrlanda, 2011.

[21] G. Alfons, **Handbuch für die Berufsvercodung**, Mannheim, 2011.

[22] K. Drasch, B. Matthes, M. Munz, W. Paulus, M. A. Valentin, **Arbeiten und Lernen im Wandel Teil V: Die Codierung der offenen Angaben zur beruflichen Tätigkeit, Ausbildung und Branche**, Nuremberg, 2012.

[23] Y. Jung, J. Yoo, S. H. Myaeng, D. C. Han, "A WebBased Automated System for Industry and Occupation Coding", *Web Information Systems Engineering - WISE 2008*, Lecture Notes in Computer Science, 5175, 443–457, 2008.

[24] F. G. Conrad, "Using Expert Systems To Model And Improve Survey Classification Processes", **Survey Measurement and Process Quality**, John Wiley & Sons, New York, 393–414, 1997.

[25] J. Hartmann, G. Schütz, **Die Klassifizierung der Berufe und der Wirtschaftszweige im Sozio-oekonomischen Panel**, Munich, 2002.

[26] K. Chu, C. Poirier, "Machine learning documentation initiative (Canada)", **Workshop on the Modernisation of Statistical Production,** İsviçre, 2015.

[27] A. Bethmann, M. Schierholz, K. Wenzig, M. Zielonka, "Automatic Coding of Occupations Using Machine Learning Algorithms for Occupation Coding in Several German Panel Surveys", **Beyond traditional survey taking. Adapting to a changing world,** Kanada, 2014.

[28] M. Belloni, A. Brugiavini, E. Meschi, K. Tijdens, "Measuring and detecting errors in occupational coding: an analysis of share data", *Journal of Official Statistics*, 32(4), 917-945, 2016.

[29] Internet: M. Beck, F. Dumpert, J. Feuerhake, Machine Learning in Official Statistics, https://arxiv.org/abs/1812.10422v1, 21.04.2019.

[30] S. By De Matteis, D. Jarvis, H. Young, A. Young, N. Allen, J. Potts, A. Darnton, L. Rushton, P. Cullinan, "Occupational self coding and automatic recording (OSCAR): an innovative validated web-based tool to collect lifetime job histories in large population", *Scandinavian Journal of Work, Environment & Health*, 43(2), 181-186,2017.

[31] A. Haslinger. Automatic Coding and Text Processing using N-grams. In Conference of European Statisticians. Statistical Standards and Studies – No. 48. Statistical Data Editing, Volume No. 2, Methods and Techniques, pages 199-209. UNO, New York and Geneva, 1997.

[32] J. Hilden, J.D.F. Habbema, B. Bjerregaard, "The measurement of performance in probabilistic diagnosis, I. The problem, descriptive tools, and measures based on classification matrices", *Methods of information in medicine*, 17, 217-226, 1978.

[33] J. Hilden, J.D.F. Habbema, B. Bjerregaard, "The measurement of performance in probabilistic diagnosis, II. Trustworthiness of the exact values of the diagnostic probabilities", *Methods of information in medicine*, 17, 227- 237, 1978.

[34] J. Hilden, J.D.F. Habbema, B. Bjerregaard, "The measurement of performance in probabilistic diagnosis, III. Methods based on continuous 54 functions of the diagnostic probabilities", *Methods of information in medicine*, 17, 238-246, 1978.

[35] Hanehalkı Bütçe İstatistikleri Mikro Veri Seti CD., Türkiye İstatistik Kurumu, Ankara, 2017.

[36] Internet: Hanehalkı Bütçe Anketinin Kapsamı, Yöntemi, Tanım ve Kavramları Hakkında Genel Açıklamalar, http://tuik.gov.tr/HbGetir.do?id=27840&tb_id=7, 12.04.2019.

[37] Internet: Household Budget Surveys in the EU, Methodology and recommendations for harmonisation, https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-BF-03-003?inheritRedirect=true, 12.04.2019.

[38] Internet: Classification of Individual Consumption According to Purpose (COICOP) 2018, https://unstats.un.org/unsd/classifications/unsdclassifications/COICOP_2018_-_pre-edited_white_cover_version_-_2018-12-26.pdf, 12.04.2019.

[39] Internet: Household Budget Survey (HBS), https://ec.europa.eu/eurostat/web/household-budget-surveys/policy-context, 12.04.2019.