TOBB EKONOMİ VE TEKNOLOJİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

HASTALIK SALGINLARININ İNTERNET ERİŞİM VE ARAMA VERİSİ KULLANILARAK TAHMİNİ

YÜKSEK LİSANS TEZİ

Batuhan BARDAK

Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı: Yrd. Doç. Dr. Mehmet TAN

AĞUSTOS 2016

Fen Bilimleri Enstitüsü Onayı

.....
Prof. Dr. Osman EROĞUL
Müdür

Bu tezin Yüksek Lisans derecesinin tüm gereksinimlerini sağladığını onaylarım.

.....
Doç. Dr. Oğuz ERGİN
Anabilimdalı Başkan Vekili

TOBB ETÜ, Fen Bilimleri Enstitüsü'nün 141111034 numaralı Yüksek Lisans öğrencisi **Batuhan BARDAK** 'nın ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı **"HASTALIK SALGINLARININ İNTERNET ERİŞİM VE ARAMA VERİSİ KULLANILARAK TAHMİNİ"** başlıklı tezi 10.08.2016 tarihinde aşağıda imzaları olan jüri tarafından kabul edilmiştir.

Tez Danışmanı: **Yrd. Doç. Dr. Mehmet TAN** .....
TOBB Ekonomi ve Teknoloji Üniversitesi

Jüri Üyeleri: **Doç. Dr. Tolga CAN (Başkan)** .....
Orta Doğu Teknik Üniversitesi

Yrd. Doç. Dr. Ahmet Murat ÖZBAYOĞLU.....
TOBB Ekonomi ve Teknoloji Üniversitesi

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, alıntı yapılan kaynaklara eksiksiz atıf yapıldığını, referansların tam olarak belirtildiğini ve ayrıca bu tezin TOBB ETÜ Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırlandığını bildiririm.

Batuhan BARDAK

ÖZET

Yüksek Lisans Tezi

HASTALIK SALGINLARININ İNTERNET ERİŞİM VE ARAMA VERİSİ KULLANILARAK TAHMİNİ

Batuhan BARDAK

TOBB Ekonomi ve Teknoloji Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı: Yrd. Doç. Dr. Mehmet TAN

Tarih: AĞUSTOS 2016

Hastalıkların hangi nedenden dolayı ortaya çıktığı ve önceden tahmin edilmesi insan sağlığı için çok önemli bir konudur. Son yıllarda teknolojinın hızla gelişmesi ve internetin yoğun biçimde kullanılmasıyla ortaya büyük miktarda veri çıkmıştır. Bu verilerden mantıklı sonuçlar çıkarmaya çalışan veri bilimciler, insanların hastalıklarla alakalı internet ortamına bıraktıkları izlerle, hastane verileri arasında ilişki aramaya başlamışlardır. Yapılan çalışma sonuçları göstermiştir ki insanların internet aramaları ile hastaneye gitmeleri arasında önemli bir ilişki mevcuttur. Tespit edilen bu ilişki kullanılarak, çeşitli hastalık salgınları tahmin edilmeye başlanmıştır.

Bu tezde temel olarak iki amaç ortaya konmuştur. Birincisi, internet arama ve erişim sıklığı verisi ile hastalık salgınlarını tahmin etmektir. İkinci amaç ise semptom olarak benzerlik gösteren hastalıkların birbiri arasındaki ilişkini saptamak ve bu ilişkinin hastalık salgınları tahmin etmekte önemi olup olmadığını incelemektir.

Yapılan ilk çalışmada Vikipedi, Google Flu Trends ve bu veri kümelerinin birleşimiyle oluşturulan modeller ile Amerika Birleşik Devletleri'ndeki grip hastalığı salgını tahmin etmeye çalışılmıştır. Elde edilen sonuçlara göre grip hastalığı salgınını tahmin etmede gayet başarılı modeller oluşturulmuştur. İlk çalışmadan alınan umut verici skorlar sayesinde ikinci çalışmada ilk çalışma genişletilmiştir. Gerçekleştirilen ikinci çalışmada ise Vikipedi ve Google Flu Trends servislerinin yanı sıra Google Trends servisinden de yararlanılmıştır. Ayrıca bu çalışmada, sadece grip hastalığı için değil, grip hastalığı ile semptom olarak benzer olabileceği düşünülen başka hastalık salgınları da tahmin edilmeye çalışılmıştır. Bu çalışmadaki bir diğer amaç ise, çoklu-iş öğrenme

yönteminden faydalanarak benzer hastalıklara ait veri kümelerinin beraber kullanılmasının hastalık salgınlarını tahmin etmedeki etkisini gözlemlemek olmuştur. Elde edilen sonuçlar ise önerilen yöntemlerin başarılı ve tutarlı olduğunu ortaya koymaktadır.

Anahtar Kelimeler: Salgın tahmini, Regresyon analizi, Makine öğrenmesi, Çoklu-iş öğrenimi, İnternet servisleri, Veri birleştirme

ABSTRACT

Master of Science

FORECASTING DISEASE OUTBREAKS BY USING INTERNET ACCESS AND SEARCH DATA

Batuhan BARDAK

TOBB University of Economics and Technology
Institute of Natural and Applied Sciences
Department of Computer Engineering

Supervisor: Asst. Prof. Mehmet TAN

Date: AUGUST 2016

Tracking source of the disease and the forecasting the disease outbreaks are vital topic for human life. In recent years, with the rapid development of technology and wide usage of the internet, the amount of data that can be collected to extract meaningful information from the data with data scientists. Data scientists began to seek a relationship between the internet search data and hospital reports. Results of the studies have shown that, there is a relationship between people with internet searches, and their visits to hospitals. Using this relationship, significant amount of research is introduced to predict disease outbreaks.

The two objectives outlined in this thesis as the basis. The first objective is, forecasting the disease outbreaks by using frequency data. Second one is to determine the relationship of diseases that share similar symptoms and decide whether this relationship is of importance on forecasting disease outbreaks.

Firstly, in this study, Wikipedia, Google Flu Trends and models that are created by the combination of these data sets to predict influenza in the United States of America was tried. According to the results, the models are quite successful in predicting the flu epidemic were created. In the second study, in addition to Wikipedia and Google Flu Trends, Google Trends was also used. In addition, this study does not only cover the influenza disease, but also tries to forecast other disease which have similar symptoms with influenza. Moreover, in this study, the relationship between disease and improvements of the usage of similar disease data sets together were examined. The proposed method reveals the success of the resulting outputs.

Keywords: w

TEŞEKKÜR

Yüksek lisans eğitimim ve tez çalışmalarım boyunca desteğini ve yardımını esirgemeyen, bana sevdiğim bir alanda araştırma imkanı sağlayan değerli hocam Yrd. Doç. Dr. Mehmet TAN 'a sonsuz teşekkürlerimi sunarım.

Bu süreçte kıymetli tecrübelerinden faydalandığım TOBB Ekonomi ve Teknoloji Üniversitesi bölümünün değerli öğretim üyelerine, sunduğu güzel çalışma ortamı ve burs imkanı ile beni destekleyen değerli TOBB Ekonomi ve Teknoloji Üniversitesi ailesine minnettarım.

Birlikte çalışmaktan mutluluk duyduğum asistan arkadaşlarıma, özellikle de bu zorlu yüksek lisans sürecini başarmayı kolaylaştıran oda arkadaşlarıma teşekkür ederim.

Son ve en önemli olarak da, hayatımın her döneminde beni destekleyen, bana her aşamada yol gösteren ve her zaman yanımda olan aileme gönülden teşekkürlerimi sunarım.

İÇİNDEKİLER

ŞEKİL LİSTESİ

Sayfa

ÇİZELGE LİSTESİ

Sayfa

KISALTMALAR

ABD : Amerika Birleşik Devletleri
CDC : Centers for Disease Control and Prevention
ETL : Extract-Transform-Load
JSON : JavaScript Object Notation
GC : Google Correlate
GFT : Google Flu Trends
GT : Google Trends
LASSO : Least Absolute Shrinkage and Selection Operator
MTL : Multi-task Learning
OLS : Ordinary Least Square
STL : Single-task Learning
RAM : Random Access Memory
ILI : Influenza Like Illness

SEMBOL LİSTESİ

Bu çalışmada kullanılmış olan simgeler açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler	**Açıklama**
r	Pearson korelasyonu
d	Hastalık ismi
w_0	Sabit terim (intercept)
w_1	Eğim katsayısı (slope)
x	Bağımsız değişken
ε	Hata terimi
α	Düzenleştirme parametresi
ρ	L1, L2 norm dengeleyicisi
z_i	Gerçek değerin normalize edilmiş hali
r^2	r kare skoru
MSE	Mean Squared Error

1. GİRİŞ

Her sene yüzbinlerce insan, grip, kızamık, boğmaca ve lyme gibi bir çok hastalığa yakalanmaktadır. Bu hastalıklar insan sağlığını tehdit etmekle beraber hastanelerdeki tedavi süreçleri, ilaç tedarikleri ve benzeri nedenlerden dolayı da ülke ekonomilerine ciddi anlamda zarar vermektedir. Özellikle grip hastalığı dünyada en yaygın görülen hastalık türlerindendir. Dünya genelinde, grip hastalığı nedeniyle her yıl 3 ile 5 milyon arasında vaka ve 250 ile 500 bin arasında ölüm gerçekleşmektedir [45]. Amerika Birleşik Devletleri(ABD)'nde ise her yıl grip nedeniyle meydana gelen ölüm sayısı 3 ile 49 bin kişi arasında değişmektedir. Verilen istatistiki bilgilerden de görülebileceği üzere, gelişmiş ülkeler de dahil bütün dünya ülkeleri, birtakım hastalıkların ağır sonuçlarına katlanmaktadır. Bu nedenle birçok ülke, grip başta olmak üzere, hastalıkların önlenmesi ve tedavisi için milyonlarca lira para harcamaktadır [28].

Bu hastalıkların insan sağlığına ve ülke ekonomilerine verdiği zararı en aza indirmek için hastalık salgınlarının ortaya çıkışının önceden tahmin edilip gerekli önlemlerin alınması, toplumun bu konuda mümkün olan en erken zamanda uyarılması, hastane ve kliniklerde gerekli ilaç ve diğer lojistik düzenlemelerin yapılması önemli bir konu haline gelmiştir. Son zamanlarda hastalık salgınlarının erken tespiti ve gözlemi alanında birçok çalışma gerçekleştirilmeye başlanmıştır. Bu çalışmalar, daha eski tarihlerde hastane raporlarının analizi ile gerçekleştirilirken, günümüzde internetin yüksek oranda kullanılması sayesinde, Twitter ve Facebook üzerinden paylaşılan gönderilerle, Vikipedi ve Google gibi internet sitelerinin erişim ve arama verileriyle yapılabilmektedir.

Bu tez çalışmasında, hastalık salgınlarının internet erişim ve arama verileri ile tahmini, hastalık verileri arasındaki ilişki ve hastalık verilerinin beraber kullanılmasının hastalık salgını tahminine olan etkisi üzerinde çalışılmıştır. Oluşturulan bütün modeller ve yapılan deneyler, Amerika Birleşik Devletleri için gerçekleştirilmiştir. Bunun nedeni, ABD'de internetin yüksek oranda kullanımı nedeni ile insanların internet ortamında bıraktıkları izin fazla olması, Amerika Birleşik Devletleri Hastalık Kontrol ve Korunma Merkezleri(CDC) kurumunun bahsedilen hastalıklar dahil diğer birçok hastalığın ülke genelinde ne kadar sık görüldüğünü gözlemlemekte olması, raporları paylaşması ve literatürdeki çalışmaların çoğunlukla bu bölge için olmasıdır. Hem internet kullanımının giderek yaygınlaşması, hem de bazı kurumların hastane ve klinik verilerini paylaşarak gerçekte kaç kişinin hangi nedenlerle hastaneye gittiği bilgisinin bilinmesi üzerine, veri bilimi ile uğraşan insanlar bu verileri kullanarak hastalık salgınları için erken uyarı ve gözlem uygulamaları geliştirmeye başlamışlardır.

Bu geliştirilen uygulamalar sonucunda oluşturulabilecek sistemlerle beraber hastalık salgınları önceden tespit edilerek ve gerekli önleyici tedbirler alınarak oluşabilecek vaka sayısı en aza indirilmeye çalışılacaktır. Bunu başarmanın en önemli ve giderek popüler hale gelmeye başlayan yöntemi ise, insanların internet üzerinden yaptıkları aramaları ve internet ortamına bıraktıkları izleri kullanmaktan geçmektedir. Bilindiği üzere teknolojinin hızla gelişmesi ve internetin günümüzde yaygın olarak kullanılmasıyla beraber sosyal medya ve internet aramaları da insanların davranış biçimleri ve genel trendlerin tespiti için sıkça kullanılmaya başlanmıştır.

İnternet ortamındaki veriler ile trend tespiti konusundaki çalışmalarda 2 ana veri kaynağı kullanılmaktadır. Bunlar sosyal medya ve çeşitli internet sitelerinin erişim ve aranma sıklığı veri kaynaklarıdır. Sosyal medya veri kaynaklarına Facebook, Twitter, Instagram gibi uygulamalar örnek gösterilebilir. Bu uygulamalardan veri toplama ve toplanan verinin anlamlandırılması için duygu analizi, doğal dil işleme(NLP) gibi teknikler gerekmektedir. Bunun sebebi atılan bir tweet'in ya da paylaşılan bir Facebook gönderisinin hangi konuyla ilgili olursa olsun içeriğinin olumlu ya da olumsuz yönde olup olmadığı bilinmesi gerekliliğidir. Öte yandan, Vikipedi, Google gibi popüler internet siteleri üzerinde yapılan sorguların sıklığı incelenerek çeşitli konular hakkında trend tahmini yapılabilmektedir. Bu tez çalışmasında Vikipedi ve Google internet sitelerinin sağladığı servisler, erişim ve aranma sıklığı verileri kullanılmıştır.

Vikipedi, birçok insanın aradığı bilgiye ulaşmak için kullandığı ve artık bir standart haline gelmiş yeni nesil internet ansiklopedisidir. Bu tezin yazıldığı zamanda, yaklaşık olarak 5 milyon İngilizce makale Vikipedi bünyesinde mevcut olup, Vikipedi internette en çok aranan 7. internet sitesi konumundaydı [2]. Vikipedi üzerinde bulunan makalelere kaç kere tıklandığı, erişildiği, ile ilgili istatistiksel bilginin ,Vikipedi tarafından, paylaşılmasıyla beraber Vikipedi sadece son kullanıcılar için değil aynı zamanda da veri bilimi ile uğraşan insanlar için önemli bir veri kaynağı olmaya başlamıştır.

İnternette arama yapmak ve bilgiye ulaşmak için kullanılan bir diğer büyük internet sitesi Google'dır. Google son kullanıcılar için mevcut en büyük arama moturu olmasıyla beraber çeşitli alt servisleri mevcuttur. Bu tezde kullanılan Google servisleri şunlardır: Google Flu Trends(GFT), Google Trends(GT) ve Google Correlate(GC). Google Flu Trends, 25 ülkede grip hastalığının seviyesini ölçme, tahmin etme ve gözlemleme amacıyla oluşturulmuş bir internet servisidir. GFT, Google üzerinde yapılan sorguları inceleyerek ülke ve şehir bazlı grip aktivite tahmini yapmaktadır. Bizim çalışmalarımızda da grip hastalığı ile ilgili Google'a ait veriler GFT aracılığı ile toplanmıştır.

Tez çalışması kapsamında kullanılan bir diğer veri servisi Google Trends(GT)'dir. Google Trends, seçilen bir anahtar kelimenin/cümlenin google sorgularını baz alarak belli bir lokasyon ve zaman bilgisine göre ne kadar sık arandığını paylaşan servistir. Tez çalışmalarında, grip hastalığı haricinde incelenen diğer hastalıklara ait veriler, Google Trends veri servisi aracılığı ile toplanmıştır.

Google Correlate [27] tez çalışmalarında kullanılan son Google servisidir. Google Correlate, girilen bir anahtar kelimenin, Google sorgularına göre benzer arama sıklığı gösteren anahtar kelimeleri döner. Grip haricindeki diğer hastalıklar için Google Trends'den veri indirmeden önce, hangi anahtar kelimelerin verilerini indirmemiz gerektiğini belirlerken, hastalıkların isimleri Google Correlate servisine girilerek hastalıklar ile alakalı benzer anahtar kelimeler bu servis aracılığı ile bulunmuştur.

Toplanan bu internet servisi verileri ve Amerika Birleşik Devletleri Hastalık Kontrol ve Korunma Merkezleri'nin paylaşmış olduğu gerçek hastane verileri üzerine çeşitli makine öğrenmesi algoritmaları uygulayarak hastalık salgınlarının önceden tahmini ve hastalıkların birbiri arasındaki ilişki tez kapsamında incelenmiştir.

Bu tez çalışması şu şekilde düzenlenmiştir. Bölüm 1'de, tez çalışması hakkında genel bilgiler ve kullanılan servis ve yöntemler anlatılmıştır. Bölüm 2'de, literatürdeki benzer çalışmalar ele alınmış ve açıklanmıştır. Bölüm 3'de bu çalışmada kullanılan algoritmalar olan lineer regresyon ve çoklu-iş öğrenme algoritmalarına değinilmiş, oluşturulan modelin nasıl doğrulandığı ve düzenleştirme (regularization) yöntemleri anlatılmıştır. Bölüm 4'de, çalışmamız esnasında kullandığımız verilerin nasıl toplandığı, toplanma işleminden sonra veri üzerine yapılan ön işlemler(preprocessing) ve öznitelik çıkarma(feature extraction) yöntemlerinden bahsedilmiştir. Bölüm 5'de, hastalıkların tahmini ve ilişkisi için oluşturulan yöntemlerin sonuçları paylaşılmış ve birbirleri ile karşılaştırılmıştır. Ayrıca bu sonuçlar üzerinden modellerin güçlü ve zayıf yönlerinin nedenleri hakkında açıklamalar yapılmıştır. Tez çalışmasının sonuncu ve 6. Bölümünde, yapılan deneyler ve çalışmaların sonuçları açıklanmış, gelecek çalışmalardan bahsedilip, tez sonlandırılmıştır.

2. İLGİLİ ÇALIŞMALAR

Tez çalışması kapsamında CDC, Vikipedi, Google Trends, Google Flu Trends servisleri kullanılarak veri toplanmıştır. Bu bölümde, tez çalışmasında kullandığımız veri sağlayıcılarını kullanarak benzer çalışmalar yapan literatürdeki diğer çalışmalar incelenecektir.

2.1 Geleneksel Veri Sağlayan Servisler ile Yapılan Çalışmalar

Geleneksel veri sağlayacıları ifadesi ile anlatılmak istenen, verinin internet ortamı ya da sosyal medya gibi ortamlardan elde edilmesi yerine doğrudan kurum, hastane, klinik ve benzeri ortamlarda oluşturulan raporlardan elde edilmesidir. Örneğin, Amerika Birleşik Devletleri'nde, CDC kuruluşu düzenli bir şekilde ülke genelinde çeşitli hastalıklar nedeni ile klinik ve hastanelere gelen kişi sayısını ve diğer benzeri istatistikleri paylaşmaktadır. Bu ve benzeri kurumların verileri kullanarak grip ve diğer hastalıklar için önleyici ve erken tahmin edici sistemler kurulmaya çalışılmaktadır.

Hastane verileri haricinde, acil servis hattının aranma sıklığı, okul ve iş hayatındaki rapor alınma verileri gibi veri kümeleri ile de çalışmalar yapılmıştır [19]. Bir başka çalışmada ise sıcaklık ve nem verilerinin hastalıklar ile ilişkisi incelenmiştir [38].

Bu tür veri kümeleri ile yapılan çalışmaların önemli bir avantajı ve dezavantajı vardır. En önemli avantajı, eldeki verilerin doğruluğunun kesin olmasıdır çünkü veriler sadece kişiler gerçekten hastaneye gittiğinde ya da acil servis hattını aradığında toplanmaktadır. Bu avantajın beraberinde getirdiği dezavantaj ise verilerin toplanması ve yayınlanması arasında 1-2 haftalık gecikmenin mevcut olmasıdır. Bu gecikme de tahmin işleminin zamanında yapılmasına engel olmaktadır.

2.2 Vikipedi Servisi Kullanılarak Yapılan Çalışmalar

Vikipedi, 2001 yılında kurulan, internet ansiklopedisi alanında en popüler internet sitesidir. Milyonlarca kullanıcının erişim sağladığı Vikipedi, makalelerine olan erişim sıklık verisini paylaşıma açmıştır. Bu veriler birçok çalışma alanında kullanılmaya başlamıştır. En sık yapılan çalışmalar genellikle popüler haber başlıklarını ve olayları tespit etme alanındadır. Bu çalışmalar haricinde, sinema filmlerinde gişe başarımı tahmini [25], borsa hisseleri [26] gibi ekonomiye yönelik çalışmalar da mevcuttur. Sağlık alanında da bir takım çalışmalar gerçekleştirilmiştir. Örneğin, dengue ateşi [3], [6], kanser [21], [34], ilaç bilgileri [9] gibi alanlarda farklı çalışmalar yapılmıştır.

Literatürde, Vikipedi erişim verilerini kullanarak hastalıkların tahmini konusunda araştırma yapan benzer çalışmalar mevcuttur. Bu alanda örnek verilebilecek ilk çalışma Tausczik ve arkadaşlarına ait olan, H1N1 virüsüyle alakalı Vikipedi makalelerinin trafiğini inceleyen çalışmadır [39]. Bir diğer çalışmada Aitken ve arkadaşları, ilaç satışları ile sağlıkla alakalı 5000 Vikipedi makalesi trafiği arasında bir korelasyon tespit etmiştir [1]. McIver & Brownstein'ın çalışmasında Vikipedi erişim verileri kullanılarak Amerika içindeki grip seviyesini tahmin etmek için LASSO regresyon kullanılarak Poisson modeli oluşturulmuştur [23]. Tez çalışmamıza en çok benzeyen iki makale ise [16] ve [14]'dur. [16] çalışmasında Vikipedi erişim verisinden yararlanılmış ve Amerika'daki grip seviyesini ölçmek için sezonsal SEIR(Susceptible, Exposed, Infected, Resistant) modeli oluşturulmuştur. Ayrıca modelin sürekli olarak geliştirilebilmesi için Kalman filtresi kullanılmıştır. [14] çalışmasında ise değişik ülkelerdeki farklı hastalıklar için Vikipedi makale erişim sıklık verileri ve resmi hastane verileri toplanmıştır. Vikipedi makalelerinin erişim sıklığı, o dile ait bütün makalelere olan erişim sıklığına göre normalize edilmiş ve hastane verileri ile korelasyonu incelenerek, korelasyonu en yüksek olan hastalıklarla ilgili makaleler seçilmiştir. Toplamda 14 değişik ülke-hastalık kombinasyonu üzerinde tahmin işlemi gerçekleştirilmiş olup, bunlardan 8 tane model hastalık salgınını yüksek skor ile tahmin ederken, 6 model de ise makalede açıklanan bir takım sebeplerden dolayı hastalık salgınlarını başarılı bir şekilde tahmin edememiştir.

2.3 Google Servisleri ile Yapılan Çalışmalar

Bu alt başlıkta Google Trends ve Google Flu Trends servislerinden yararlanılarak yapılan literatürdeki bir takım çalışmalara değinilmiştir. Google Trends servisinin ortaya çıkış amacı Google üzerinde yapılan sorguların lokasyon, dil ve zamana göre ortaya çıkan trendlerini görmektir. Bu servisin ürettiği verileri kullanarak bir çok alanda tahmin işlemi yapılmıştır. Örneğin, [22] çalışmasında petrol fiyatları, [13] araç satış miktarı ve [8] çalışmasında da işsizlik şikayetleri ile ilgili tahminlerde bulunulmuştur. Sağlık alanında da Google Trends servisini kullanarak çalışmalar yapan araştırmacılar bulunmaktadır. [37] lyme hastalığı, [47] listeriosis, [30] sıtma , [4] genel hastalık salgınlarının tahmini, [7] Kore'deki hastane verileri ile Google Trends servisinden gelen verilerin korelasyonunu ölçmek ve [10] de Google internet sorguları ile Norovirüs hastalığı ilişkisi üzerine çalışmalar gerçekleştirilmiştir.

Google Flu Trends servisi ise Google'ın sadece grip hastalığını tahmin etmeye ve gözlemlemeye yönelik oluşturduğu internet servisidir. Bu servis doğrudan Google sorgularını inceleyerek çeşitli ülkelerdeki grip seviyesini ölçmeye çalışmaktadır. Literatürde bu servis kullanılarak grip hastalığını tahmin etmeye yönelik çalışmalar gerçekleştirilmiştir. Örneğin, [32], [31], [11] çalışmalarında Google Flu Trends servis verileri kullanılarak grip hastalığı tahmin edilmeye çalışılmıştır. [48] çalışmasında ise topluluk modeli(ensemble) ve veri servisi beraber kullanılarak grip hastalığı tahmin edilmeye çalışılmıştır. Veri kümesi olarak, CDC, Athenahealth, Google Trends, Twitter, FluNearYou ve Google Flu Trends kullanılarak veriler toplanmış ve deneyler yapılmıştır. Bu deneyler sırasında, Stacked Lineer Regresyon, SVM regresyon algoritmalarından yararlanılmıştır. Bir diğer çalışmada [35], Google Flu Trends servisinin kısıtlamaları göz önünde bulundurularak Google Trends ve Google Correlation servislerini kullanan

ARGO(AutoRegression with Google) modeli oluşturulmuştur. Araştırmacıların hipotezlerine göre oluşturulan model, elde edilen CDC verisini otomatik olarak toplayarak en alakalı Google sorgularını seçebilmektedir.

Bu çalışmada belirtildiği üzere Google Flu Trends servisinin bazı kısıtlamaları vardır [35]. Ayrıca 2012/2013 sezonunda Google Flu Trends servisi grip tahminlerinde gerçek sonuca göre oldukça yüksek tahminler yapmıştır [20]. Bu durumdan sonra, Google modelini geliştirmiş ve güncellemiştir.

3. KULLANILAN YÖNTEMLER

Bu bölümde çalışmalarımızda kullanılan makine öğrenmesi algoritmaları ve modelimizi eğitirken kullandığımız bazı yöntemler açıklanmıştır.

3.1 Lineer Regresyon

Makine öğrenmesinde algoritmalar genel olarak gözeticili, gözeticisiz, yarı gözeticili, pekiştirmeli, uyum ve öğrenmeyi öğrenme algoritmaları olarak ayrılmıştır. Lineer regresyon algoritmaları gözeticili algoritma grubuna ait olup, bir veya birden fazla bağımsız değişken ile bir bağımlı değişken arasındaki ilişkiyi bulmak için kullanılır.

Sadece tek bağımsız değişken(x) varsa bu yapıya basit lineer regresyon adı verilir ve aşağıdaki gibi formülize edilir:

$$h_w(x) = w_0 x_0 + w_1 x_1 \tag{3.1}$$

Burada x bağımsız(açıklayıcı) değişken, w_0 (intercept) sabit terim ve w_1(slope) eğim katsayısıdır. Eğim katsayısı(w_1), x'deki değişimin $h_w(x)$'e ne kadar etki edeceğini gösterir.

Çoklu lineer regresyon modelleri ise basit lineer regresyon modellerine göre daha kompleks yapıdadır ve m adet bağımsız değişken olduğu varsayılırsa, formülü aşağıdaki gibidir:

$$h_w(x) = w_0 x_0 + w_1 x_1 + w_2 x_2 + .. + w_m x_m \tag{3.2}$$

Her iki denklemde de $x_0 = 1$'dir ve sabit bir terimdir. Her iki formül de kapalı formda aşağıdaki gibi ifade edilebilir:

$$h_w(x) = \sum_{i=0}^{m} w_i x_i = w^T x \tag{3.3}$$

Burada T ifadesi transpoz işlemini ifade eder.

Veri kümesindeki herhangi bir i. noktayı tahmin etme işlemi aşağıdaki formül ile yapılabilir:

$$y_i = h_w(x_i) + \varepsilon_i \tag{3.4}$$

Formülde y_i tahmin etmeye çalıştığımız bağımlı değişken, $h_w(x_i)$, x verisi ile fonksiyonumuzun döndürdüğü sonuç, ε_i ise fonksiyonumuzun tahmini ile gerçekte tahmin etmeye çalıştığımız değer arasındaki hata miktarını simgeler. Bir başka deyişle hata terimi, y üzerinde etkili olan x'in dışındaki diğer faktörleri temsil eder.

Oluşturulan modellerin başarılı tahminler üretmesi için teoride $\mathbf{E}[\varepsilon_i] = 0$ olması beklenmektedir. Burada $\mathbf{E}$ beklenen değer ifadesini simgelemektedir.

Formül 3.5 ise Artıkların Kare Toplamı(AKT,RSS:Residual sum of squares)'dır.

$$AKT = \sum_{i}^{N} (y_i - h_w(x_i))^2 = \sum_{i}^{N} \varepsilon_i^2 \tag{3.5}$$

Başarılı tahminler yapan bir modelde AKT değerinin düşük olması beklenmektedir. Çünkü düşük AKT değeri, modelin yaptığı tahmin değeri $\hat{y}$ ile gerçek y arasındaki farkın az olması anlamına gelmektedir. Bu aynı zamanda modelimizin veriyi ne kadar iyi açıkladığını belirtir. ATK değeri modelimizin maliyet fonksiyonudur ve bu değeri minimum yapabilmek için hipotezdeki w parametrelerinin değiştirilmesi ve ATK değerini minimum yapan w değerleri kombinasyonun bulunması gerekmektedir. Bu bir optimizasyon problemidir ve *Gradient Descent* algoritması bu problem için kullanılabilir. Bu algoritmada iteratif olarak bütün parametrelerin birinci türevi alınır ve bu parametreler güncellenir. ATK fonksiyonu üzerinde türev alma işleminin daha kolay olması için standart olarak formülün başına $1/2$ eklenir. ATK fonksiyonun yeni hali maliyet fonksiyonu olarak aşağıdaki gibi yazılabilir.

$$J(w) = \frac{1}{2} \sum_{i}^{N} (h_w(x^i) - y^i)^2 \tag{3.6}$$

Gradient Descent algoritmasında kullanmak üzere maliyet fonkisiyonumuzun türevi aşağıdaki gibidir:

$$\frac{\partial J}{\partial w_j} = -\sum_{i=1}^{n} (y^i - h_w(x^i))(x_j^i) \tag{3.7}$$

Gradient descent algoritması $J(w)$ fonksiyonunu minimum yapacak olan w değerlerini bulmaya çalışır. Algoritmanın her iterasyonundan hemen sonra aşağıdaki formüle göre ağırlık güncellemesi yapılır.

$$\Delta w_j = -\alpha \frac{\partial J}{\partial w_j} \tag{3.8}$$

Bu formüldeki α değeri algoritmanın öğrenme hızının katsayısıdır. Bu değer çok küçük seçildiği takdirde algoritma yavaş çalışabilir, yüksek seçildiği takdirde ise minimum noktasında salınım yapabilir, minimum noktasını kaçırabilir ya da minimum değere ulaşmadaki iterasyon sayısı artabilir. α değerinin seçiminde farklı yöntemler

kullanılabilir. Örneğin, α değeri sabit bir değer alınabilir veya her iterasyonda belirli miktarda azaltılabilir $\alpha(t+1) = \alpha(t)/\sqrt[2]{t}$.

$$w := w + \Delta w \tag{3.9}$$

Formül 3.9'daki işlem bütün $j = 1, 2, 3, ..., n$ değeri için yapılır. Algoritma 1'de Gradient Descent algoritması verilmiştir.

Gradient Descent algoritması literatürde *Batch GD* olarak da geçer ve bazı durumlar da problemlerle karşılaşılır. Gradient Descent algoritmasında gradient maliyeti bütün eğitim seti bazında hesaplandığından, çok büyük veri setleri kullanıldığında bu yöntem masraflı ve yavaş olabilir. Bu duruma karşı alternatif olarak *Stochastic Gradiendt Descent*(SGD) kullanılabilir. SGD her bir iterasyonda her bir örneği gezdikten sonra güncelleme işlemini gerçekleştirir.

Algoritma 1 Batch Gradient Descent Algoritması

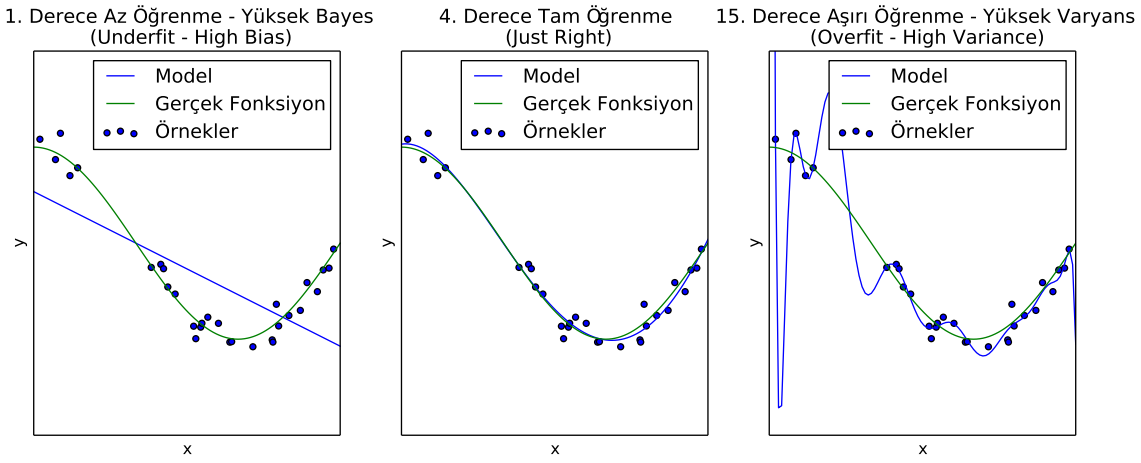
1: **for** 1 veya daha fazla iterasyon için **do**
2: **for** Her bir j ağırlığı için **do**
3: $w_j = w + \Delta w_j, (\Delta w_j = \alpha \sum_{i=1}^{n} (y^i - h_w(x^i))(x_j^i))$

3.2 Model Seçimi ve Performans Analizi

Makine öğrenmesi algoritmaları ile oluşturulan modellerin seçimi ve performans analizi önemli bir konudur. Oluşturulan modeller arasından test verisi karşısında en yüksek sonucu verecek modeli seçmek ve performans analizini doğru yapmak için mevcut veri kümesi eğitim, doğrulama ve test kümeleri olarak üç parçaya ayrılamalıdır. Bu ayırma işlemi %70 - %15 - %15 şeklinde olabileceği gibi farklı oranlarda ayrım yapmak da mümkündür. Bu ayrımın amacı, veri kümesinin büyük kısmı ile modeller oluşturup, doğrulama veri kümesi ile en iyi sonuç veren modelin seçilmesidir. Modelin başarımının güvenilir olması için modelin daha önce hiç görmediği test veri kümesi kullanılarak modelin başarımı ölçülmelidir.

Modelin, yeni veriler karşısındaki başarımının beklenenden düşük çıkması durumunda oluşturulan model veya veri kümesi üzerinde bazı problemlerin mevcut olduğu sonucuna varılabilmektedir. Bu problemlerin çözümünde ise aşağıdaki yöntemlerden bazıları kullanılabilir:

- Daha çok veri ile eğitim gerçekleştirme
- Öznitelik sayısını azaltma (örneğin: PCA algoritması)
- Daha fazla öznitelik ekleme
- Polinomsal öznitelik ekleme
- Düzenleştirme (regularization) terimini (α) azaltma
- Düzenleştirme (regularization) terimini (α) arttırma



Şekil 3.1: Doğrulama ve eğitim hatasının model karmaşıklığı ile ilişkisinin gösterimi.

Bu yöntemlerden hangisinin ne zaman kullanılması gerektiğine ise model ve veri kümesi üzerinde yapılacak olan incelemeler ile karar verilebilmektedir. Modelin beklenen sonuçtan daha kötü olduğu durumlardaki en önemli problem modelin az öğrenmiş(underfit) veya aşırı öğrenmiş(overfit) olduğu durumlardır. Şekil 3.1'de bu durumlar grafiksel olarak gösterilmiş ve açıklanmıştır.

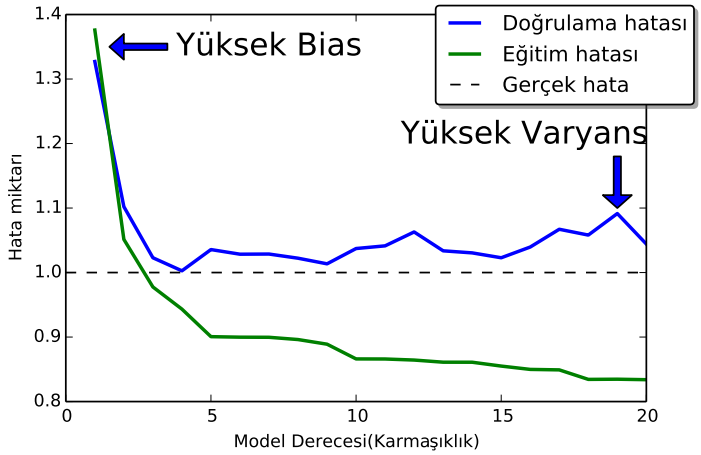
Makine öğrenmesinde az öğrenme ile aşırı öğrenme arasında bir ödünleşme mevcuttur. Oluşturulan modelin, eğitim veri kümesini çok iyi bir şekilde açıklıyor olması modelin yeni gelen verileri de çok iyi tahmin etmesi anlamına gelmemektedir. Bu duruma aşırı öğrenme denmekle beraber Şekil 3.1'de 3. grafikte bu durum gözlemlenmektedir. Modelin karmaşıklığı arttıkça, modelin eğitim veri kümesine olan uygunluğu artarken test verisi üzerindeki başarımı azalmaktadır. Bu durumun tam tersi az öğrenme problemidir. Az öğrenme durumunda oluşturulan modelin karmaşıklığı çok düşüktür ve veriyi tam olarak açıklayamamaktadır. Başarılı bir model oluşturmak için model karmaşıklığı dengelenerek Şekil 3.1'in ortasındaki grafikteki gibi bir model yaratılması gerekir.

Bu durum aynı zamanda bias-varyans ilişkisi ile ilgilidir. Bias, tahmin edilmeye çalışılan gerçek değer ile modelin tahmini değer arasındaki farkı, Varyans ise modelin eğitim kümesindeki bir değişikliğe ne kadar duyarlığı olduğunu açıklar. Şekil 3.2'de gözüktüğü gibi modelin karmaşıklığı arttıkça eğitim verisindeki hata düşmekte, doğrulama veri kümesindeki hata ise belli bir noktaya kadar düşmekte daha sonra tekrar artışa geçmektedir. Bu yüzden bias-varyans dengesi göz önünde bulundurularak, model için en uygun karmaşıklık derecesi seçilmeli ve modelin az veya aşırı öğrenme durumuna düşmesi engellenmelidir. Oluşturulan model istenilen sonuçların altında kaldığında modelin bias mi yoksa varyans problemi mi yaşadığını anlamak için Şekil 3.2'deki gibi bir grafik çıkartılabilir. Eğer modelde yüksek bias problemi varsa eğitim veri kümesinin hata miktarı yüksektir. Ayrıca eğitim ile doğrulama veri kümelerinin hata oranları birbirine yakındır. Eğer modelde yüksek varyans problemi varsa eğitim kümesinin hata oranı çok düşük olur. Bu durumda oluşan başka bir nokta ise, doğrulama veri kümesinin hata miktarının eğitim verisinin hata miktarına göre oldukça yüksek olmasıdır.

Eğitilen modellerin başarımı beklenenden düşük çıktığında yapabileceğimiz bir diğer işlem bir sonraki bölümde de detaylıca anlatılan düzenleştirme terimi α değerini azaltıp arttırmaktır.

α değeri büyük seçilirse:

- Bütün öznitelikler yüksek miktarda cezaya uğrar.
- Bu yüzden çoğu öznitelik 0 değerine yaklaşır.
- Böylece hipotezimiz de 0 değerine yaklaşır.
- Bu durum da karmaşıklığı az olan bir model elde edilir ve bu bize yüksek bias'e sahip ve az öğrenmiş bir model verir.



Şekil 3.2: Doğrulama ve eğitim verileri üzerinde değişen hata miktarı ve bias-varyans ilişkisinin gösterimi.

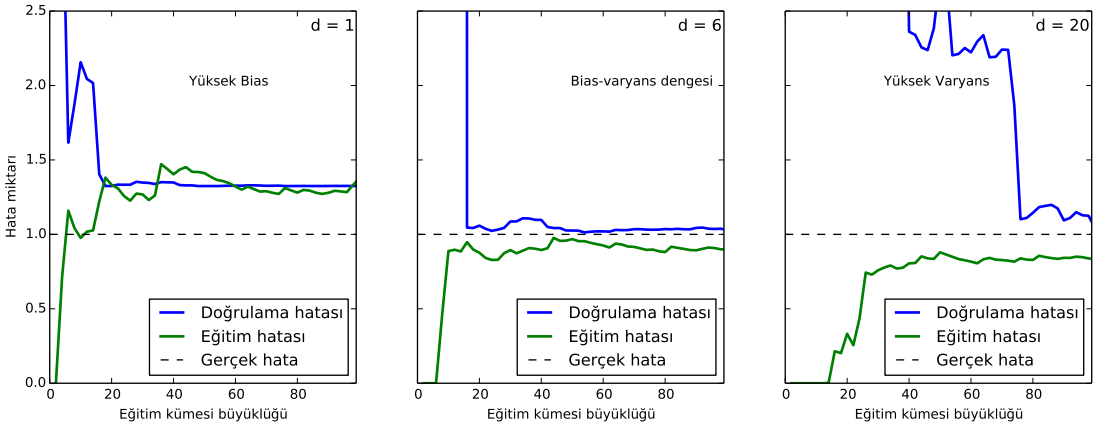
α değeri orta büyüklükte seçilirse:

- Sadece seçilen bu değerde model mantıklı sonuçlar verebilir.

α değeri küçük seçilirse:

- $\lambda = 0$
- Böylece düzenleştirme terimi 0 yapılmış olur.
- Öznitelikler hiç cezaya uğramaz ya da çok az uğrar.
- Bu durum modelin yüksek varyans yani aşırı öğrenme problemi yaşamasına neden olur..

Başarımı yükseltmek ve algoritmik doğruluğu ölçmek için öğrenme eğrileri oluşturmak da önemli bir yöntemdir. Bu yöntem sayesinde eğitilen modelin doğruluğu ve yapısı gözlemlenebilmektedir. Aşağıdaki çeşitli bias-varyans durumlarına göre oluşabilecek grafikler Şekil 3.3'de gösterilmiştir.



Şekil 3.3: Bias-varyans değişimine göre eğitim ve doğrulama veri kümesi hata ilişkisi

- Veri yüksek bias'e sahip ise Şekil 3.3'ün solundaki gibi bir grafik ortaya çıkar. Eğitim ve doğrulama hatası birbirine yakın ve yüksektir.
- Veri bias-varyans dengesine sahip ise Şekil 3.3'ün ortasındaki gibi bir grafik ortaya çıkar. Eğitim ve doğrulama hatası birbirine yakın ve düşüktür.
- Veri yüksek varyans'a sahip ise Şekil 3.3'ün sağındaki gibi bir grafik ortaya çıkar. Eğitim ve doğrulama hatası arasındaki fark yüksek ve eğitim hatası düşüktür.

Özetle, oluşturulan model beklenenden düşük performans gösteriyor ise bunun temel iki sebebi olabilir: yüksek bias (az öğrenme) ve yüksek varyans(aşırı öğrenme). Öğrenme eğrisi grafiğini çizerek, eğitim kümesi ve doğrulama kümesi üzerindeki hata miktarına bakılarak problemin nerede olduğu anlaşılmaya çalışılabilir.

Eğer problem **Yüksek Bias** ise:

- Daha fazla öznitelik eklemek
- Polinomsal öznitelikler ekleyerek modelin karmaşıklığını arttırmak
- Düzenleştirme parametresi α'yı azaltmak

Eğer problem **Yüksek Varyans** ise:

- Eğitim kümesinin boyutunu büyütmek
- Daha az öznitelik ile çalışmak
- Düzenleştirme parametresi α'yı arttırmak

yöntemleri uygulanabilir.

3.3 Düzenleştirme(Regularization)

Bölüm 3.2'de regresyon modelinin karşılaşabileceği bazı problemlerden bahsedilmiştir. Bu problemlerin nasıl çözülebileceği ile ilgili bir önceki alt başlıkta anlatılan *düzenleştirme* yöntemi bu bölümde detaylıca ele alınmıştır. Düzenleştirme yönteminin çözmeye çalıştığı temel iki sorun aşağıda listelenmiştir:

- Öznitelik sayısının, örnek sayısından çok büyük olduğu durumlar
- Modelin aşırı öğrenme problemi yaşadığı durumlar

Oluşturulmaya çalışılan başarılı bir modelin, test veri kümesindeki verileri başarılı bir şekilde tahmin etmesi ve eğitim veri kümesini ezberlememesi beklenir. Bunun için modelin ne kadar kompleks olacağını belirlemek gerekir. Modelin ne kadar kompleks olması gerektiği Bölüm 3.2 anlatıldığı gibi öğrenme eğrisi grafikleri çizilerek belirlenmeye çalışılabilir. Bir diğer yöntem ise regresyon formülüne ceza terimi ekleyerek, lineer regresyon modelinin katsayılarını cezalandırmaktır. Bu bağlamda üç düzenleştirme yöntemi ele alınmıştır. Bunlar:

- Ridge: L2 Norm'undan yararlanır.
- LASSO: L1 Norm'undan yararlanır.
- Elastic Net: Hem L1 hem de L2 Norm'unun avantajlarından yararlanır.

3.3.1 Ridge

Ridge [29], L2 normunu kullanan bir düzenleştirme yöntemidir. Ridge regresyon, OLS'nin minimize etmeye çalıştığı amaç fonksiyonuna, katsayıların L2 normunu ekleyerek katsayıların cezalandırılması sağlar.

OLS'nin minimize etmeye çalıştığı fonksiyonunun gösterimi Formül 3.10'da gösterilmiştir.

$$\min_{w} ||Xw - y||_2^2 = \min_{w} (\sum_{i}^{n} (y_i - \hat{y_i})^2) \tag{3.10}$$

Ridge tarafından eklenen katsayıların L2 normu ve bu normun katsayısı(α) aşağıda belirtilmiştir:

$$\alpha ||w||_2^2 = \alpha \sum_{i} w_i^2 \tag{3.11}$$

Buradaki α değeri L2 normunun katsayısı olup cezalandırma işleminin ağırlığını belirler. Eğer α değeri 0 olarak seçilir ise Ridge regresyonu, OLS regresyon modeli ile aynı hale gelir. α değeri arttıkça model katsayıları düzleşmeye başlar.

O halde Ridge regresyonun optimizasyon amacı Formül 3.12'deki gibi ifade edilebilir.

$$\min_w ||Xw - y||_2^2 + \alpha||w||_2^2 = \min_w (\sum_i^n (y_i - \hat{y}_i)^2) + \alpha \sum_i w_i^2 \tag{3.12}$$

Formüldeki α değeri 0 iken Ridge regresyonu OLS ile aynı görevi görmektedir. α değerinin artmasıyla beraber katsayılar üzerinde cezalandırma artmaktadır fakat katsayıları bir sonraki bölümde anlatılan LASSO düzenleştiricisi gibi 0 yapmamaktadır.

3.3.2 LASSO

Düzenleştirme tekniklerinden popüler bir diğer yöntem LASSO'dur. LASSO [40] yönteminin Ridge yönteminden farkı, cezalandırma yönteminde farklı bir norm olan L1 normunu kullanmasıdır.

L1 normu ve bu normun katsayısı aşağıda Formül 3.13'de belirtilmiştir:

$$\alpha * ||w||_1 = \alpha \sum_i |w_i| \tag{3.13}$$

O halde LASSO düzenleştirici yönteminin optimizasyon hedefi aşağıda belirtilen Formül 3.14'de gösterilmiştir:

$$\min_w ||Xw - y||_2^2 + \alpha||w||_1 = \min_w (\sum_i^n (y_i - \hat{y}_i)^2) + \alpha \sum_i |w_i| \tag{3.14}$$

Ridge ile LASSO yöntemleri arasında katsayıları nasıl cezalandırdıkları ile ilgili bir fark vardır. Ridge katsayıların değerini azaltmaya yönelik cezalandırma işlemi yaparken LASSO bu katsayıları 0'lamaya ve seyrek katsayılardan oluşan bir regresyon modeli yaratmaya çalışır. LASSO düzenleştirmesinde α değeri büyüdükçe katsayılar 0 değerini almaya başlar.

3.3.3 Elastic Net

Elastic Net düzenleştirme yöntemi, LASSO yönteminin bir takım limitasyonlarını engellemek için ortaya çıkmıştır. LASSO, birbiri ile korelasyonu yüksek parametrelerden sadece birini seçerek diğerlerini gözardı eder. Bu yüzden Elastic Net, LASSO düzenleştiriciyle beraber Ridge düzenleştiricisini de kullanılır. Böylece L1 ve L2 normunu kullanarak Ridge ve LASSO yöntemlerinin avantajlarını birleştirir [49]. ρ parametresi L1 normu ile L2 normu arasındaki ilişkiyi dengeler. Formül 3.15, Elastic Net yönteminin amaç fonksiyonunu vermektedir.

$$\min_w ||Xw - y||_2^2 + \alpha\rho||w||_1 + \alpha(1-\rho)||w||_2^2 \tag{3.15}$$

$$= \min_w (\sum_i^n (y_i - \hat{y}_i)^2) + \alpha\rho \sum_i |w_i| + \alpha(1-\rho) \sum_i w_i^2$$

3.4 Çoklu-iş Öğrenme(Multi-task Learning)

Makine öğrenmesi alanındaki yöntemler genelde aynı anda sadece bir işi öğrenmeye çalışırlar. Bu durum tek-iş öğrenme(STL-Single-task learning) olarak adlandırılabilir. Çoklu-iş öğrenme yapısı ise birbiriyle alakalı problemleri beraber çözer [5]. Çoklu-iş öğrenme yapısı öğrenilmesi gereken işleri beraber ele almaktadır. Beraber ele alınan bu işler arasında bir ilişki varsa öğrenme işlemi, işleri tek tek öğrenmekten daha avantajlı bir hale gelebilmektedir [36]. Çoklu-iş öğrenme yapısı tez çalışması kapsamında, hastalık salgın tahminlerinin başarımı arttırmada, hastalıkların birbiri arasındaki ilişkiyi incelemede kullanılmıştır. [12] [49] çalışmalarında çoklu-iş öğrenmeyle ile ilgili daha detaylı bilgilendirme mevcuttur.

Çalışmamızda çoklu-iş öğrenme yapısı önceki bölümde anlatılan Elastic net düzenleştiricisi kullanılmıştır.

Çoklu-iş öğrenmesi ile Elastic net düzenleştiricisi beraber kullanıldığında bu yapının minimize etmeye çalıştığı hedef zarar fonksiyonu Formül 3.16'da gösterilmiştir.

$$\min_{w} ||Xw - y||_F^2 + \alpha\rho||w||_{21} + \alpha(1-\rho)||w||_F^2 \tag{3.16}$$

Bu formülde ρ ifadesi L1 ile L2 norm dengesini ayarlayan düzenleştirme parametresidir ve bu parametre için en uygun değer *grid search* yöntemi ile bir çok seçenek denenerek bulunmuştur. $||W||_F$ ifadesi Frobenius normu [44] ifade eder. ve $||W||_{21}$ ifadesi $L21$ normu aşağıdaki gibidir.

$$||W||_{21} = \sum_i \sqrt{\sum_j w_{ij}^2} \tag{3.17}$$

4. VERİ TOPLAMA ve VERİYİ İŞLEME

Bu bölümde çalışmalarımızda kullanılan veri kümelerinin hangi servislerden toplandığı, bu servislerin detayları, toplanan verilerin kullanılmadan önce hangi işlemlerden geçirildiği açıklanmıştır. Çalışmalarımızda 4 ayrı veri sağlayıcısından yararlanılmıştır.

Kullanılan veri kümelerinin alındığı servisler aşağıda listelenmiştir.

- Amerika Birleşik Devletleri Hastalık Kontrol ve Korunma Merkezleri (Centers for Disease Control and Prevention - CDC)
- Google Flu Trends
- Google Trends
- Vikipedi

Toplanan bütün veri kümeleri, 01-01-2011 ile 04-01-2014 tarihi arasındaki 158 haftalık veriyi kapsamaktadır.

4.1 Amerika Birleşik Devletleri Hastalık Kontrol ve Korunma Merkezleri

CDC, ulusal sağlık iyileştirmesi ve hastalıkların önlenmesi konusunda çalışan bir sağlık kurumudur. Bu kurum, ülke genelindeki hastaneler, klinikler ve diğer sağlık daireleri ile iş birliği içinde hareket etmektedir. Bulaşıcı, kronik, genetik ve daha birçok hastalık hakkında bilgilendirme ve hastalıkları önleyici çalışmalar yapmaktadır. Aynı zamanda ülke genelinde, belirli bir hastalık şüphesi ile hastaneye başvuran kişilerin sayısını ve istatistiksel verileri paylaşmaktadır [46]. Bu kurumun paylaştığı veriler, belli hastalık şikayeti ile hastaneye giden hasta sayısını ifade ettiği için çalışmalarımızda tahmin modelleri oluşturulurken CDC verisi baz alınarak model eğitimi gerçekleştirilmiş ve bu kurumun paylaşmış olduğu raporlardaki hastaneye başvuran kişilerin sayısı tahmin edilmeye çalışılmıştır. Grip hastalığı ile alakalı veriler bu bağlantıdan[1] , diğer hastalıklarla alakalı veriler ise bu bağlantıdan[2] indirilmiştir. Verilen bağlantılardaki bazı parametreler değiştirilerek veri kümeleri otomatik olarak indirilebilmektedir. Bahsedilen bağlantıda YY parametresi yılı, WW parametresi haftayı temsil etmektedir. Örneğin, YY= 60, WW = 01 ifadesi 2011 yılının ilk haftasını göstermektedir. Bu yaklaşımla veriler kolaylıkla toplanmıştır.

[1] `gis.cdc.gov/grasp/fluview/fluportaldashboard.html`

[2] `www.cdc.gov/mmwr/preview/mmwrhtml/mmYYWWmd.htm?s-cid=mmYYWWmd`

4.2 Google Flu Trends

Google Flu Trends(GFT), 25 ülkedeki grip seviyesini tahmin etmek ve gözlemlemek için, Google tarafından 2008 yılında duyurulmuş bir internet servisidir [15]. Bu tahmin ve gözlem işlemi, Google'da yapılan arama sorguları baz alınarak aşağıda açıklanmış olan Formül 4.1'e göre yapılmaktadır.

$$logit(P) = \beta_0 + \beta_1 \times logit(Q) + \varepsilon \tag{4.1}$$

Bu formüldeki P, grip nedeniyle hastaneye giden kişi oranını, Q ise Google üzerinden grip ile alakalı yapılan arama sorguların sıklığını temsil eder. β_0 sabit terim, β_1 eğim katsayı ve ε hata terimidir. Her ülke için seçilen milyonlarca Google sorgusu o ülkenin kendi sağlık kurum verileri ile kıyaslanarak en alakalı, korelasyonu yüksek, 45 sorgu seçilerek bir model oluşturulur. Örneğin, ABD için CDC kurumu verileri baz alınarak sorgu seçimi yapılır. Bu modeller ile ülkelerdeki haftalık grip seviyesi tahmin edilmeye çalışılmıştır.

Google Flu Trends servisinden indirilen bu veri kümesi içerisinde ABD'deki her eyalet için grip tahmin sonuçları mevcuttur. Google Flu Trends servisine ait veri GFT'nin resmi internet adresinden [42] indirildikten sonra verinin doğrudan kullanılması yerine aşağıdaki gibi bir ön işleme(öznitelik seçimi) tabii tutulmuştur.

- İndirilen Google Flu Trends veri kümesi içerisindeki her bir ABD eyalet verisi ile CDC verisinin Pearson korelasyonu r hesaplanmıştır.
- Hesaplanan $|r|$ değerleri büyükten küçüğe sıralanıp en yüksek korelasyon değerine sahip m tane eyalet seçilmiştir.

Buradaki m sayısı yapılan deneyler sonucu toplam 53 eyalet arasından 30 olarak belirlenmiştir. İndirilen Google Flu Trends verisi içerisindeki 30 eyalet verisi yukarıda anlatılan yöntem ile seçilip yeni bir veri kümesi oluşturulmuştur.

Google Flu Trends, 20 Ağustos 2015 tarihine kadar yaptığı tahminleri paylaşıma açmış bu tarihten sonra ise tahmin sonuçlarını sadece resmi internet siteleri aracılığı ile talep doğrultusunda kişi ya da kurumlarla paylaşacağını duyurmuştur.

4.3 Google Trends

Google Trends, Google tarafından, kullanıcıların Google üzerinden yaptıkları sorgu sıklıkları temel alınarak geliştirilen bir internet servisidir. Herhangi bir kelime veya cümlenin dünyanın neresinde, hangi dilde ne sıklıkta arandığını paylaşmaktadır[41]. Google Trend'in kendi resmi internet sitesinde aranma hacmi öğrenilmek istenen kelimeler sorgulanarak belirli tarih aralığındaki aranma sıklığı indirilebilmektedir. İndirilen veri [0-100] değerleri aralığına normalize edilmiş bir şekilde inmektedir. Üzerinde çalıştığımız 5 hastalıktan grip hastalığı hariç diğer dört hastalık için Vikipedi ile beraber Google Trend verileri de kullanılmıştır. Google'ın, grip haricindeki diğer hastalıklara özel Google Flu Trends gibi servisi olmadığından dolayı, diğer hastalıklar için

Google sorgularına bağlı sonuçlar manuel olarak Google Trend servisinden Google Correlate servisi yardımı ile indirilmiştir. Grip hastalığı haricindeki diğer hastalıklar için, hastalığın ismi Google Correlate servisinde aratılarak, hastalığın ismi ile benzer arama sıklığına sahip kelime ve cümleler tespit edilmeye çalışılmıştır. Hem buradan bulunan, hem de bizim hastalıkla alakalı olabileceğini düşündüğümüz anahtar terimlerin Google Trends üzerinden Google'daki aranma sıklık bilgileri indirilerek her bir hastalık için veri kümeleri oluşturulmuştur.

4.4 Vikipedi

Vikipedi, günümüzde birçok kişinin de bildiği üzere oldukça popüler bir internet ansiklopedisidir. Sitenin yüksek hacimli trafiği sebebiyle, hangi makalenin ne zaman kaç kez okunduğu bilgisi, kullanıcı davranışlarını ve genel trendi tespit etme gibi konularda kullanılmaya başlanmıştır. Bahsedilen bu veriler, 2007 yılından başlamak üzere saatlik veri olarak bu adreste [3] paylaşılmaktadır.

Vikipedi tarafından paylaşılan makale erişim sıklık bilgilerini içeren dosyaların boyutunun büyük olması ve bazı ön işlemlerden geçirme zorunluluğundan dolayı alternatif bazı internet siteleri ortaya çıkmıştır. Örneğin, [4] sitesi, Vikipedi tarafından paylaşılan saatlik veriyi toplayarak 24 saatlik, günlük, veri haline getirmekte ve JSON (Javascript Object Notation) formatı halinde paylaşıma sunmaktadır. Çalışmamızda kullanacağımız Vikipedi veri kümeleri bu internet sitesinden indirilmiştir.

Yapmış olduğumuz çalışmalarda 01-01-2011 ile 04-01-2014 tarihleri arasındaki 5 ayrı hastalık için seçilen Vikipedi makalelerinin erişim sıklık verileri toplanmıştır.

Bu hastalıklar:

- Grip (Influenza)
- Listeriosis (Listeriosis)
- Lyme (Lyme)
- Sıtma (Malaria)
- Boğmaca (Pertussis)

Grip haricinde çalışmalarda kullanılacak olan diğer dört hastalık aşağıdaki yöntem ile seçilmiştir:

- Grip hastalığına semptom olarak benzeyebilecek hastalıklar internet üzerinden araştırılmıştır.
- Araştırma sonucunda grip ile semptom olarak benzeyebileceğini düşündüğümüz 13 hastalık seçilmiştir.
- Seçilen 13 hastalığın CDC verisinin, CDC grip verisi ile Pearson korelasyonu r değeri hesaplanmıştır.

[3] `dumps.wikimedia.org/other/pagecounts-raw/`

[4] `stats.grok.se`

- En yüksek |r| değerine sahip 4 hastalık seçilmiştir.

Çalışmada grip haricinde 4 hastalıkla çalışılıp, 5. hastalığın seçilmeme sebebi $|r|$ değerinin düşük olmasından kaynaklanmaktadır.

Seçilen bu hastalıklar ile ilgili hangi makalelerin erişim sıklık verisinin indirileceği, indirildikten sonra verinin hangi işlemlerden geçirileceği aşağıda açıklanmıştır. Çalışmalarda kullanılan 5 hastalıktan grip hastalığı için farklı diğer hastalıklar için farklı yöntem uygulanmıştır.

Grip hastalığı için başka bir çalışmada[14] belirlenen, griple alakalı olabilecek 53 İngilizce Vikipedi makalesi ele alınmıştır. Bahsedilen çalışmadaki yazarların açıklamalarına göre bu makaleler grip ve sağlık alanında yetkin kişiler tarafından seçilmiştir. Seçilen makaleler bizim tarafımızdan da incelendiğinde, makalelerin grip hastalığı ile ilgili olduğu kanısına varılmıştır. Oluşturulacak olan modelde başarımı yükseltmek adına seçilen bu 53 makale doğrudan kullanılmamış olup aşağıdaki yöntem ile seçilmiştir:

- Seçilen 53 makalenin her birinin, CDC verisi ile Pearson korelasyonu r değeri hesaplanmıştır.
- Hesaplanan $|r|$ değerleri büyükten küçüğe doğru sıralanmış ve en büyük değerli k tanesi seçilmiştir.

Buradaki k değeri yapılan farklı deneylere göre 10 olarak belirlenmiştir ve griple alakalı 53 makale içinden, CDC grip verisi ile korelasyonu en yüksek 10 makale seçilmiştir.

Diğer dört hastalık için ise bu işlem biraz daha farklı bir şekilde gerçekleştirilmiştir. Bunun nedeni, diğer hastalıklara grip kadar sık rastlanılmıyor olmasındandır. Bu hastalıkların daha az rastlanır olması, internetteki arama sorgularının verilerinin de hacmini düşüreceğinden bu veri kümelerine farklı işlemler uygulanmıştır.

Diğer her hastalık d için, $d \in$ {Listeriosis, Lyme, Sıtma, Boğmaca}:

- d ile alakalı 10 tane Vikipedi makalesi seçilmiştir.
- Google Correlate servisi kullanılarak 10 tane d ile alakalı anahtar kelime seçilmiştir.
- Bu 10 anahtar kelimenin Google'da aranma hacmini bulabilmek için Google Trends'den aranma sıklıkları sayıları(0-100 aralığına normalize edilmiş) indirilmiştir.
- 10 Google Trends anahtar kelime verisini m ve 10 Vikipedi makale erişim sıklığı verisini n olarak kabul edersek, her bir anahtar kelime ve makale erişim sıklığı verisinin CDC verisi ile Pearson korelasyonu hesaplanmış, en yüksek değeri veren 10 anahtar kelime veya makale seçilmiştir.

4 hastalık için eğitim verisi toplamda Google Trends ve Vikipedi verileri birleştirilerek, $m+n = 10$ olacak şekilde seçilmiştir. 10 sayısı yapılan deneyler sonucu en iyi sonucu

Çizelge 4.1: Her hastalığa ait toplanan veri kümeleri. Koyu renkli işaretlemeler veri kümesinin modelde kullanıldığını temsil etmektedir.

	GFT	GT (**m**)	Vikipedi (**n**)	GT+Vikipedi	CDC
Grip	**30 eyalet**	x	**10 makale**	x	**158h**
Listerosis	x	10 kelime	10 makale	**m + n = 10**	**158h**
Lyme	x	10 kelime	10 makale	**m + n = 10**	**158h**
Sıtma	x	10 kelime	10 makale	**m + n = 10**	**158h**
Boğmaca	x	10 kelime	10 makale	**m + n = 10**	**158h**

verdiği için seçilmiştir. Çizelge 4.1'de her hastalık için toplanan bütün veri kümeleri topluca gösterilmiştir. Koyu ile işaretli olan veri kümeleri modellerin eğitimi sırasında kullanılan veri kümelerini olduğunu göstermektedir.

4.5 Normalizasyon ve ETL Süreci

Bu alt başlıkta, toplanan ve oluşturulan veri kümelerinin normalizasyon yönteminden ve genel ETL(Extract-Transform-Load) sürecinden bahsedilmiştir. Normalizasyon işlemi Formül4.2'ye göre yapılmıştır.

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{4.2}$$

Bu formülde z değeri gerçek değerin normalize edilmiş halini, x değeri ise orjinal örneği temsil etmektedir. i ise veri kümesinin i. elemanını temsil etmektedir.

Tahmin edilmeye çalışılan bütün hastalık verileri CDC kurumunun internet sitesinden indirilmiştir. İndirilen bu verilerin değer aralığı birbirinden farklı türde olduğu için, grip haricindeki dört hastalığın CDC verisi, CDC Grip verisi baz alınarak Formül 4.3'e göre normalize edilmiştir.

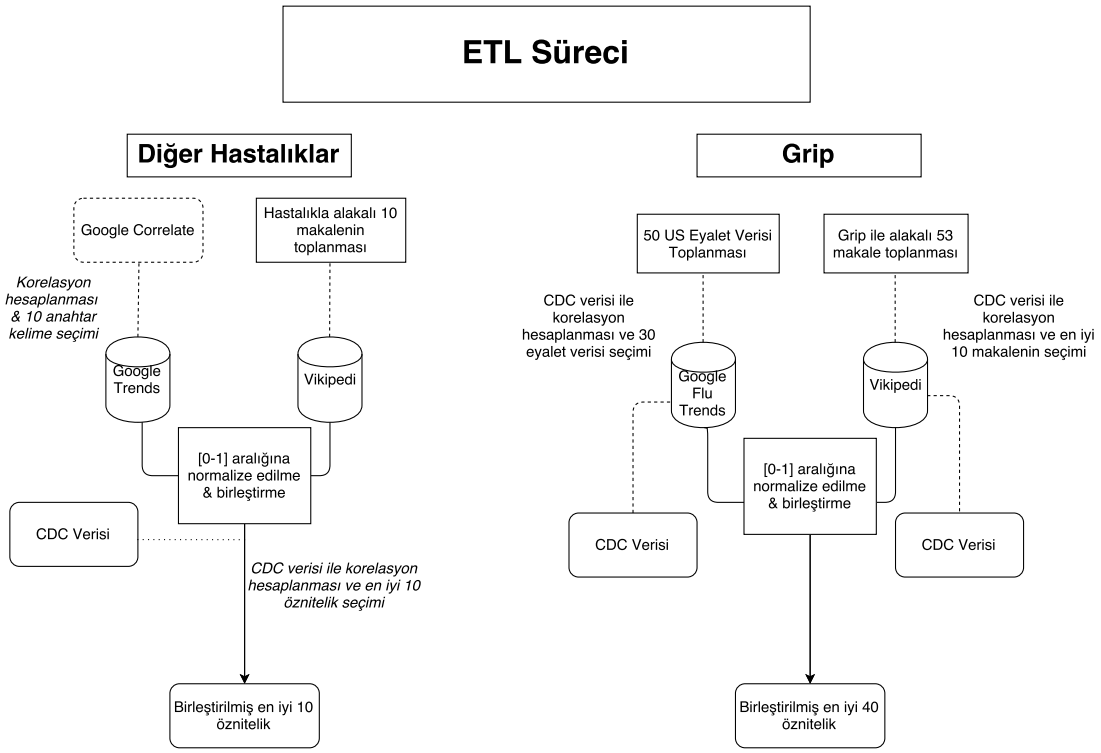
$$z_i = a + \frac{x_i - \min(x)(b-a)}{\max(x) - \min(x)} \tag{4.3}$$

Formülde a ve b değerleri CDC Grip hastalığı verisinin minimum ve maksimum değerlerini temsil etmektedir. Bu bölümde şimdiye kadar anlatılan bütün bu veri toplama, veriyi dönüştürme, veriyi modele vermek için hazır hale getirme işlemleri Şekil 4.1'de toplu halde gösterilmiştir.

Yapılan bu veri toplama, veri ön işleme ve normalizasyon süreci sonucunda oluşan ve çalışmalarımızda kullanacağımız veri kümeleri aşağıda listelenmiştir.

- Grip hastalığı için CDC kurumuna ait veri kümesi (158 Haftalık).
- Diğer dört hastalık(Listeriosis, Lyme, Sıtma, Boğmaca) için CDC kurumuna ait, CDC grip hastalığı verisine göre normalize edilmiş veri kümesi (158 Haftalık - Grip Hastane Verisine göre normalize edilmiştir).

- Grip ile alakalı 10 Vikipedi makalesinden oluşan Veri Kümesi (161 Haftalık - Oluşturulan modellerde veri kümesi kaydırıldığı için 3 haftalık fazla veri toplanmıştır, [0-1] aralığına normalize edilmiştir).
- Google Flu Trends'den yapılan öznitelik seçimi işlemi sonucu 30 eyaletlik veri kümesi (161 Haftalık - Oluşturulan modellerde veri kümesi kaydırıldığı için 3 haftalık fazla veri toplanmıştır, [0-1] aralığına normalize edilmiştir).
- Diğer dört hastalık için Vikipedi makalesi m, Google Trends anahtar sözcük aranma hacmi n olmak üzere ve $m+n=10$ olacak şekilde oluşturulan veri kümesi (161 Haftalık - Oluşturulan modellerde veri kümesi kaydırıldığı için 3 haftalık fazla veri toplanmıştır, [0-1] aralığına normalize edilmiştir).



Şekil 4.1: Veri toplama, Dönüştürme and Hazır hale getirme(ETL) süreci şeması

5. DENEYSEL SONUÇLAR

Bu bölümde, Bölüm 3'de açıklanan Lineer regresyon, Ridge, LASSO, Elastic Net düzenleştiricileri ve çoklu-iş öğrenmeyle(multi-task learning) beraber Bölüm 4'de anlatılan CDC, Vikipedi, Google Trends ve Google Flu Trends servislerinden elde edilerek oluşturulan veri kümeleri ile gerçekleştirilen 2 ayrı çalışmanın sonuçlarına yer verilmiştir. İlk çalışmamızda, Vikipedi, Google Flu Trends ve bu veri kümelerinin birleştirilmesiyle oluşturulan yeni veri kümesi ile ABD'de yaşayan insanların ne kadarının grip hastalığı şüphesiyle hastane ve kliniklere gideceği tahmin edilmeye çalışılmıştır. İkinci çalışmamızda ise ilk çalışmamız genişletilerek sadece grip hastalığı değil başka hastalıklar da tahmin edilmeye çalışılmıştır. Bu işlemi gerçekleştirirken Vikipedi ve Google Flu Trends servislerine ek olarak Google Trends servisi kullanılmıştır. Bu çalışmada hastalıklar tahmin edilmeye çalışılırken hastalıklara ait veri kümelerini beraber kullanarak hastalıklar arasındaki olası ilişkinin bu tahmin işleminde olumlu bir etkisi olup olmadığı da araştırılmıştır.

Tez kapsamındaki ilk çalışma Bölüm5.2'de, ikinci çalışma ise Bölüm5.3'de detaylı bir şekilde açıklanmıştır.

5.1 Ayarlar

Yapılan çalışmalar Python dilinde kodlanmıştır. Veri analizi için Pandas [24], makine öğrenmesi algoritması için scikit-learn [33], güçlü dizi işlemleri için Numpy, matematiksel işlemler için Scipy [18], görselleştirme ve grafikler için Matplotlib [17] ve internetten veri çekmek için BeautifulSoup ve Urllib [43] kütüphanelerinden yararlanılmıştır.

Model sonuçları doğrulanması için 10-katlı çapraz doğrulama yöntemi kullanılmıştır. Ayrıca model parametreleri, eğitim kümesi üzerinde *grid search* yöntemi ile iyileştirilmiştir. Bütün yapılan deneyler, 3.2GHz işlemci ve 16 GB RAM ve Linux işletim sistemine sahip bilgisayar üzerinde gerçekleştirilmiştir.

5.2 Vikipedi ve Google Flu Trends Verilerinin Birleştirilmesiyle Grip Salgını Tahmini

Tez kapsamındaki ilk çalışmada Bölüm 4'de anlatıldığı gibi grip hastalığının tahmini için Vikipedi ve Google Flu Trends servisleri aracılığı ile oluşturulan veri kümeleri kullanılmıştır. Ayrıca bu veri kümeleri birleştirilerek de yeni veri kümesi elde edilmiş ve oluşturulan bu veri kümesi ile deneyler yapılmıştır. Yapılan deneylerin sonuçları incelenerek bu veri birleştirme işleminin grip hastalığı salgınını tahmin etmedeki etkisi incelenmiştir.

Özetle, bu işlemi gerçekleştirmek için yapılan deneyler esnasında 3 ayrı veri kümesi ile işlem yapılmıştır.

Bu veri kümeleri:

- Grip hastalığı ile alakalı 10 Vikipedi makalesinin erişim sıklığı verisini içeren veri kümesi.
- ABD'ye ait 30 eyaletin GFT veri kümesi.
- Yukarıda belirtilen iki veri kümesinin birleştirilmesiyle oluşturulan yeni veri kümesi(Vikipedi + GFT).

Oluşturulan bu veri kümeleri Bölüm 3'de anlatılan makine öğrenmesi algoritmalarıyla eğitilerek çeşitli modeller oluşturulmuştur. Öncelikle oluşturulan modellerin başarımının hangi parameterelere göre ve nasıl ölçüldüğü daha sonra ise model sonuçları açıklanmıştır. Model başarımında dikkate alınan 2 metrik şunlardır:

- Modelin ne kadar önceden yüksek başarımla tahmin yapabildiği (*Offset*)
- Modelin doğruluğu (r^2 *ve MSE*)

İlk metrik, oluşturulan modelin hastaneye grip şüphesi ile gidecek olan insanların sayısını ne kadar önceden tahmin edebildiğini(bu metrike **offset** adı verilmiştir), ikinci metrik ise bu tahminlerin ne kadar doğru olduğunu açıklamaktadır.

5.2.1 Offset kavramı

Offset, kısaca hastaneye giden insanların, hastaneye gittikleri gün ile hastaneye gitmeden önce hastalıklarını internette araştırdıkları gün arasındaki zaman dilimini ifade etmektedir. Oluşturulan modellerin en başarılı şekilde kaç gün önceden tahmin yapabildiğini bulabilmek adına elimizdeki veri kümeleri kaydırılarak birçok *offset* değeri için modeller oluşturulmuştur.

Bu kaydırma işlemi Vikipedi ve Google servislerinden toplanan veri kümeleri için farklı biçimde yapılmıştır. Bunun nedeni, Vikipedi veri kümesinin günlük, Google Flu Trends veri kümesinin ise haftalık veri türünden oluşmasıdır. Ayrıca tez kapsamında

yapılan ilk çalışmada -21, +7 günleri arasında ikinci çalışmada ise -21, 0 günleri arasında kaydırma işlemi yapılmıştır. İlk çalışmada -21 ile +7 gün arasında deneme yapılma sebebi, literatürdeki benzer bir çalışmanın [14] bu offset değerlerini kullanıyor olmasıdır. İkinci çalışmada ise 0 ile +7 gün arasında veri kaydırmanın gerekli olmadığına karar vererek bu offset değerleri için deney yapılmamıştır.

Bu sebeple ilk çalışmamızda Vikipedi veri kümesi -21 gün ile +7 gün arasında, GFT veri kümesi ise -3 hafta ile 1 hafta arasında kaydırılarak modellemeler tekrarlanmıştır. Bu kaydırma işlemleri sonucunda 29 Vikipedi, 5 GFT, 145 (29x5) tane de Vikipedi+GFT veri kümesi ile eğitilmiş model oluşmuştur.

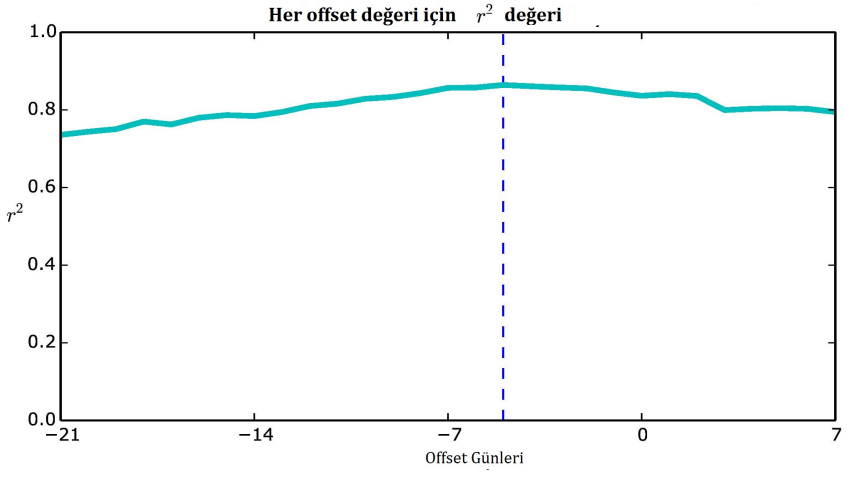
İkinci çalışmamızda Vikipedi ve GFT/GT veri kümeleri, ilk çalışmanın son modelinde olduğu gibi beraber kullanılmıştır. Vikipedi veri kümesi -21 gün ile 0 gün arasında, GFT ve GT veri kümesi ise -3 hafta ile 0 hafta arasında kaydırılmıştır. Toplamda 22 Vikipedi x 4 GFT/GT veri kümesi ile 88 tane veri kümesi kombinasyonu ile model oluşturulmuştur.

İlk çalışma için modellerin offset kombinasyonlarını örneklemek gerekirse:

- **Vikipedi**: $-21W, -20W, -19W.., +7W$: 29 model
- **GFT**: $-3GFT, -2GFT, -1GFT, 0GFT, +1GFT$: 5 model
- **Vikipedi+GFT**: $(-21W, -3GFT), (-21W, -2GFT)..(7W, +1GFT)$: 29x5=145 model

İkinci çalışma için modelleri örneklemek gerekirse:

- **Vikipedi+GFT/GT**: $(-21W, -3GFT/GT), (-21W, -2GFT/GT)..$ $..(0W, +0GFT/GT)$: 22x4=88 model



Şekil 5.1: Vikipedi veri setini -21, +7 gün kaydırarak oluşturulmuş farklı modellerin skorları

Oluşturulan bu modellerin başarımını ölçmek için ise r^2 ve MSE(mean squared error) metrikleri kullanılmıştır.

r^2 metriği, regresyon analizinde elde edilen denklemin bağımlı değişkenini ne kadar iyi açıkladığını anlatan metriktir. r^2 metriğinin formülü, Formül 5.1'de gösterilmiştir.

$$r^2 = 1 - \frac{\sum_{i=0}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n}(y_i - \bar{y})^2} \tag{5.1}$$

burada $\hat{y}_i$, i. örnek için tahmin değeridir ve,

$$\bar{y} = \frac{1}{n}\sum_{i=0}^{n} y_i \tag{5.2}$$

İkinci metrik olan ortalama kareler hatası(MSE) ise modelin tahmin ettiği ile gerçek değerler arasındaki farkı temsil eder ve Formül 5.3'deki gibi hesaplanır.

$$MSE = \frac{1}{n}\sum_{i}^{n}(y_i - \hat{y}_i)^2 \tag{5.3}$$

Bu düzenlemeler altında Vikipedi veri kümesinin -21, +7 günleri arasında kaydırılarak elde edilen 29 modelinin r^2 değişkenine bağlı olarak değişen grafiği Şekil 5.1'de gösterilmiştir. Bu grafiğe göre en yüksek r^2 skoru -5. günde gözlemlenmiştir. Bu sonuç, sadece Vikipedi veri kümesi kullanıldığı takdirde grip hastalığı nedeniyle hastanaye gidecek kişi sayısının en yüksek doğrulukta 5 gün önceden tahmin edilebildiğini göstermektedir. Bu kaydırma işlemi(*offset*) diğer veri kümeleri üzerinde de denendiğinde en iyi *offset* değerleri Çizelge 5.1'deki gibi ortaya çıkmıştır.

Bu modeller oluşturulurken kullanılan makine öğrenmesi algoritmaları ise Bölüm 3'de anlatılan OLS, Ridge regresyon, LASSO regresyon, Elastic net regresyon olmuştur. Çalışmalarda kullanılan veri kümelerinin yapısı nedeniyle Ridge, LASSO, Elastic net gibi düzenleştirme yöntemlerinden yararlanılmıştır. Veri kümelerinin yapısı ifadesi ile antılmak istenen, verinin 158 örnek, ve 40 öznitelik içermesidir. Bu sayılara bakıldığında modeli eğitmek için gereken örnek sayımızın az, öznitelik sayımızın ise çok olduğu görülmektedir. Bu durum Bölüm 3.2'de de detaylıca anlatıldığı üzere modelin aşırı öğrenmesine sebep olabileceğinden düzenleştirme algoritmaları denenmiş ve başarılı sonuçlar verdiği görülmüştür. Oluşturulan modele ait sonuçları paylaşılırken sadece en iyi sonuç veren algoritmalara ait sonuçlar paylaşılmıştır.

Elimizdeki üç ayrı veri kümesinin -21 ile +7 günleri arasında kaydırılması sonucu eğitilen modellerin en başarılı tahmin sonuçlarına ulaştıkları günler(offset değerleri) Çizelge 5.1'de paylaşılmıştır.

Çizelge 5.1: En iyi offset zamanları

Model	**Gün**	**Hafta**
Vikipedi	-5	Saatlik Veri
Google Flu Trends	Haftalık Veri	-1
Vikipedi & GFT	-5	-1

5.2.2 Vikipedi veri kümesi ile oluşturulan model

Bu çalışma kapsamında ilk olarak Vikipedi servisinden toplanan veri kümesi ile modeller oluşturulmuştur. Daha önce bahsedildiği üzere, Vikipedi veri kümesi günlük veri içerdiği için -21,+7 günleri arasında kaydırılmış ve her ayrı *offset* değeri için toplamda 29 model oluşturulmuştur. Oluşturulan bu modellere ait en başarılı sonuçlar Çizelge 5.2'de paylaşılmıştır.

Çizelgenin ilk satırında OLS, ikinci satırında ise Ridge regresyon skorları test verisi üzerinde çapraz-doğrulama yöntemi ile hesaplanmıştır. Çizelgenin son satırındaki skor ise eğitim verisi üzerinden hesaplanmış olup modelin veriye ne kadar uyduğunu göstermektedir. Vikipedi veri kümesi kullanılarak oluşturulan model sonuçları ile CDC zaman serisi verisinin ilişkisi Şekil 5.2'de gösterilmiştir.

Çizelge 5.2: Vikipedi modeli için en iyi r^2 skorları

OLS (Test Verisi ile)	0.86
Ridge Regresyon (Test Verisi ile)	0.85
OLS (Eğitim Verisi ile)	0.91

5.2.3 Google Flu Trends veri kümesi ile oluşturulan model

Oluşturulan bu modelde 30 ABD eyalet verisini içeren Google Flu Trends veri kümesi kullanılmıştır. Veri türü haftalık olduğu için veri kümesi -3 hafta ile +1 hafta arasında kaydırılarak 5 adet model oluşturulmuştur. Bölüm 5.2.2'de oluşturulan Vikipedi modeli ile aynı algoritma ve ayarlar kullanılmış olup model sonuçları Vikipedi modeline kıyasla az farkla da olsa daha başarılı çıkmıştır. Çizelge 5.3'de GFT modelinin sonuçları paylaşılmıştır. Bir önceki modelde olduğu gibi çizelgenin birinci ve ikinci satırlarındaki sonuçlar test, son satırdaki sonuç ise eğitim verisinin sonucudur. Google Flu Trends veri kümesi kullanılarak oluşturulan model sonuçları ile CDC kurumu verilerinin zaman serisi ilişkisi Şekil 5.3'de gösterilmiştir.

Çizelge 5.3: GFT modeli için en iyi r^2 skorları

OLS (Test Verisi ile)	0.86
Ridge Regresyon (Test Verisi ile)	0.86
OLS (Eğitim Verisi ile)	0.94

5.2.4 Verilerin birlikte kullanılması ile oluşturulan model

Oluşturulan son modelde, Vikipedi ve GFT veri kümeleri birleştirilerek kullanılmıştır. Bölüm 4'de anlatılan ETL sürecinden sonra ortaya çıkan 10 adet Vikipedi makale erişim sıklığı veri kümesi ile, 30 adet ABD eyalet verisinden oluşan Google Flu Trends veri kümesi birleştirilerek 40 adet özniteliğe sahip yeni bir veri kümesi oluşturulmuştur. Bölüm 2'de bahsedildiği üzere literatürde grip hastalığını tahmin etmeye yönelik

çalışmalar olmasına rağmen şu anki bilgimize göre Vikipedi ve GFT servisinden toplanan verileri birleştirerek kullanılan bir çalışmaya rastlanmamıştır. Bu model kullanılarak elde edilen çarpıcı sonuçlar, Çizelge 5.4'de paylaşılmıştır.

Sonuçlar incelendiğinde, diğer veri kümeleri ile oluşturulan modellere kıyasla her algoritmanın başırımının yaklaşık olarak %5 düzeyinde arttığı gözlemlenmektedir. Veri kümelerinin birleşimiyle oluşturduğumuz model ile CDC zaman serisi verileri ilişkisi Şekil 5.4'de gösterilmiştir.

Çizelge 5.4: Vikipedi + GFT modeli için en iyi r^2 skorları

OLS (Test Verisi ile)	0.91
Ridge Regresyon (Test Verisi ile)	0.94
OLS (Eğitim Verisi ile)	0.98

Çizelge 5.5'de, yapılan deneylerin *MSE* metriğine göre sonuçları bir arada paylaşılmıştır. Bu sonuçlardan da görüleceği üzere veri kümelerinin beraber kullanılması tahmin sonuçlarını iyileştirmiştir.

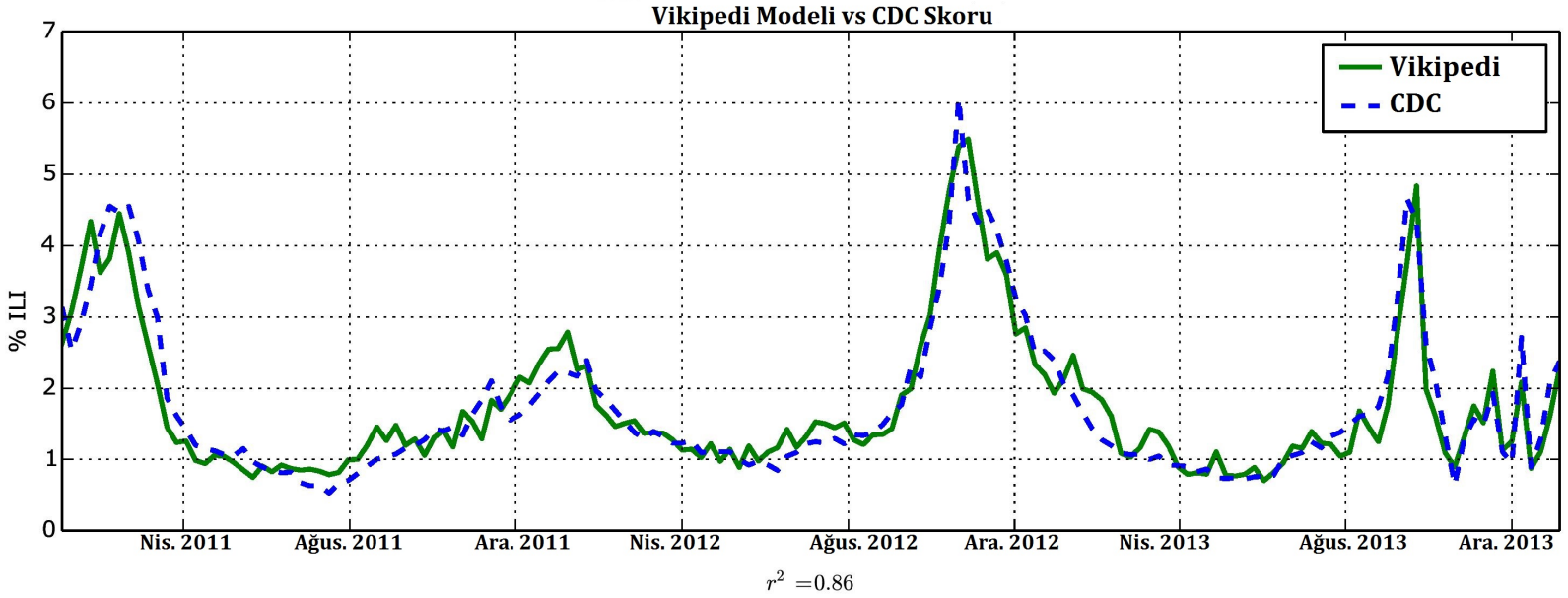
Çizelge 5.5: Deney sonuçlarının *MSE* metriği cinsinden gösterimi

Algoritma/Veri Kümesi	**Vikipedi**	**GFT**	**Vikipedi+GFT**
OLS	0.129	0.169	**0.092**
Ridge	0.142	0.145	**0.068**

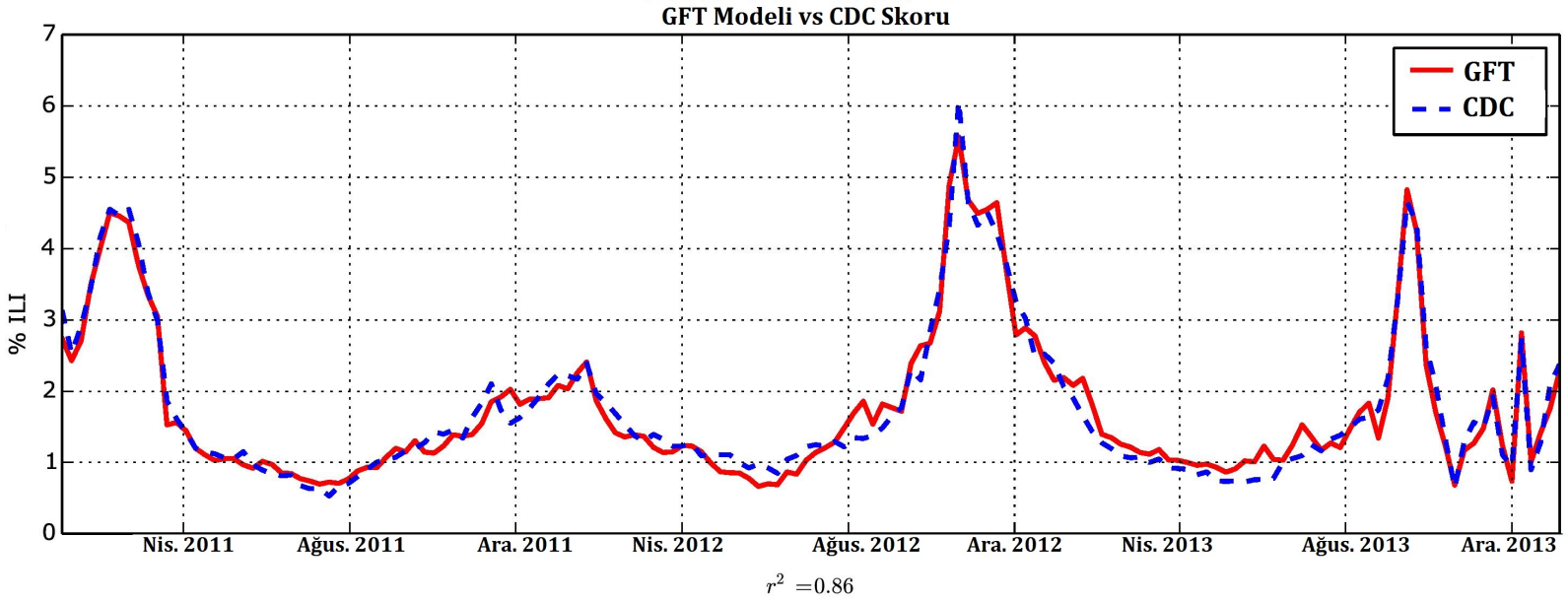
5.2.5 Tartışma

Oluşturulan modellerin sonuçları incelendiğinde, Vikipedi veri kümesi ile Google Flu Trends veri kümesinin birleştirilmesiyle oluşturulan modelin grip hastalığının tahmin edilmesinde başarımın artmasını sağladığı gözlemlenmiştir. Bu modelde en iyi sonuçlar, Vikipedi veri kümesini 5 gün, Google Flu Trends verisi de 1 hafta geriye kaydırarak bulunmuştur. Bunun anlamı, bugüne ait Google Flu Trends verisi ile 2 gün sonranın Vikipedi verisi birleştirildiğinde 7 gün sonra hastaneye grip hastalığı sebebi ile gidecek kişi sayısının en doğru şekilde tahmin edilebilir olmasıdır.

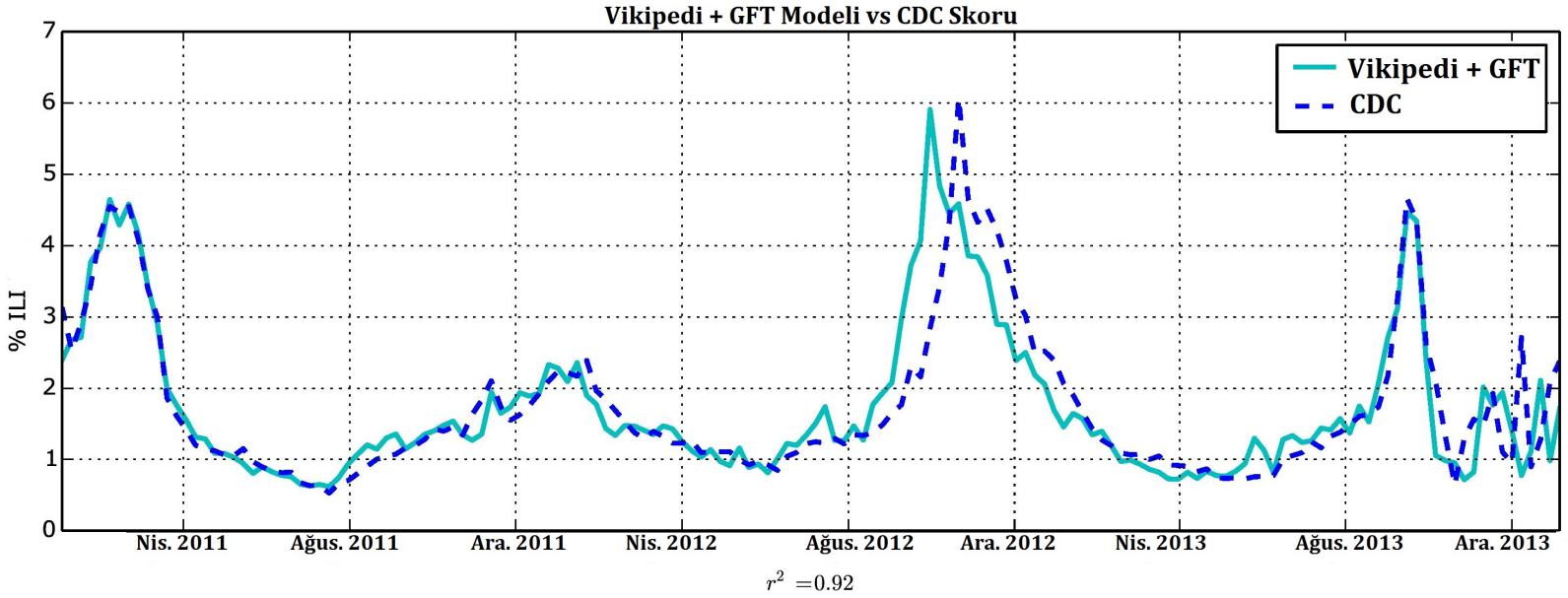
Başka bir çalışmada [14],aynı tarih aralığı için Vikipedi veri kümesi kullanarak, eğitim verisi üzerinden hesaplanan OLS r^2 skoru 0.90 olarak bulunmuştur. Bu skor, bizim Vikipedi ve Google Flu Trends verilerini birleştirerek oluşturduğumuz modelde 0.98'e yükselmiştir. Fakat bu skor, eğitim verisi üzerinde yapıldığı ve çapraz-doğrulama uygulanmadan elde edildiği için tahminden çok modelin veriyi ne kadar iyi açıkladığını ortaya koymaktadır. Bu skor türü, paylaşılan sonuç çizelgelerinde 3. satırda bulunmaktadır. Bu yaklaşım modelin yeni gelen veriyi ne kadar iyi tahmin ettiği noktasında yanıltıcı olacağından, diğer modellerde başarım ölçümü test verileri üzerinde gerçekleştirilmiştir. Bölüm 3.2'de açıklandığı ve tahmin edileceği üzere test verisi üzerindeki başarı, eğitim verisi üzerindeki başarımından düşük çıkmıştır. Eğitim verisi kümesi üzerinde gerçekleşen modellerin sonuçlarına bakıldığında Vikipedi ve GFT veri kümeleri ayrı ayrı kullanıldığında r^2 skoru 0.86 olurken, deneysel olarak oluşturulan



Şekil 5.2: Oluşturulan başarılı Vikipedi modeli ile CDC verisinin uyumunun gösterimi



Şekil 5.3: Oluşturulan başarılı GFT modeli ile CDC verisinin uyumunun gösterimi



Şekil 5.4: Oluşturulan başarılı Vikipedi+GFT modeli ile CDC verisinin uyumunun gösterimi

Vikipedi+GFT veri kümesinin r^2 skoru 0.92'ye yükselmiştir. r^2 metriği modeldeki öznitelik sayısı arttığında, artış gösterme eğiliminde olan bir metrik olduğu için aynı zamanda modellerin *MSE* değerleri de hesaplanmıştır. *MSE* skorları da incelendiğinde, Vikipedi+GFT kullanımının başarımı arttırdığı görülmüştür.

Bu çalışma yapılırken karşımıza bir takım kısıtlar çıkmıştır. Bu kısıtlar, ikinci çalışmayı gerçekleştirirken karşılaşılan kısıtlar ile ortak olduğu için bu konuya Bölüm 5.3.4'de değinilmiştir.

5.3 Hastalık Salgınlarının Veri Birleşimi ve Çoklu-iş Öğrenme Yöntemi ile Tahmin Edilmesi

Tez kapsamındaki ikinci çalışmada, ilk çalışma geliştirilerek ve genişletilerek, Amerika Birleşik Devletleri'nde yaşayan insanların grip ve grip benzeri seçilen dört hastalık nedeniyle hastaneye başvurma sayıları, internet erişim ve arama sıklık verileri ile tahmin edilmeye çalışılmıştır. Ayrıca bu çalışma kapsamında, hastalıklar arasındaki olası ilişki incelenmiş olup bu ilişkiden yararlanılarak hastalıkların verisi beraber kullanılmış ve tahmin işleminde başarı artışı yakalanmaya çalışılmıştır. Bölüm 4'de detaylı bir şekilde anlatıldığı üzere Grip hastalığı vakaalarını tahmin etmek için 10 Vikipedi makale verisi ve 30 ABD eyalet Google Flu Trends verisi birleştirilmiş ve model oluşturulmuştur. Diğer hastalıklar içinse m Vikipedia makale erişim sıklık sayısı, n Google Trends anahtar kelime sayısı olmak üzere $m+n=10$ olacak şekilde veri kümesi oluşturulmuş ve her hastalık için modeller bu veri setleri ile eğitilmiştir.

Oluşturulan üç modelin ilki Elastic Net düzenleştirmesi ile çoklu lineer regresyonu (Multiple linear regression with Elastic Net) kullanılarak eğitilmiştir. İkinci ve üçüncü model ise Elastic Net düzenleştirmesi ile çoklu-iş öğrenme yönteminden faydalanarak (Multi-task learning with Elastic Net regression) eğitilmiştir. Model başarımları Formül 5.1'de açıklanan r^2 ve 5.3'de açıklanan *MSE* metriklerine göre karşılaştırılmıştır.

Çalışmanın amaçları ve oluşturulan modeller aşağıda listelenmiştir:

- Her hastalığın meydana gelme sıklığının, hastalıkla ilişkili veri kümeleri ile tahmin etme (Tek Veri ile Hastalık Tahmini-Model 1)
- Grip hastalığı ile diğer hastalık verilerinin birleştirilmesi (ikili kombinasyon) ile bu hastalıkların meydana gelme sıklığının tahmin edilmesi ve bu iki hastalık arasındaki ilişkiyi saptama (Hastalıkların ikili kombinasyonları - Model 2)
- Bütün hastalık verilerinin beraber kullanılması ile hastalıkların meydana gelme sıklığının tahmin edilmesi (Model 3)

5.3.1 Tek hastalık verisi ile tahmin

Bu çalışma kapsamında oluşturulan ilk modelde, her hastalığın CDC kurumuna ait sonuçları, hastalıkların internetteki aranma sıklığı verileri ile tahmin edilmeye çalışılmıştır. Bu model, tez kapsamındaki ilk çalışmamızda oluşturduğumuz grip hastalığını tahmin etme modelinin diğer hastalıklar için genişletilmiş versiyonudur.

Bu bölümde aşağıdaki veri kümeleri kullanılarak her hastalık için birer model oluşturulmuştur.

- Grip veri kümesi
- Listeriosis veri kümesi
- Lyme veri kümesi
- Sıtma veri kümesi
- Boğmaca veri kümesi

Çizelge 5.6, her hastalığın CDC verilerinin kendi modeli ile tahmin edildiğindeki başarımını *MSE* metriği cinsinden göstermektedir. Bu çizelgede modellerin *MSE* skorları ile beraber offset değerleride paylaşılmıştır. Offset değerlerinin başındaki **-** işareti, geçmiş günü ifade etmektedir. Örneğin, -21W ifadesi, Vikipedi veri kümesinin 21 gün geçmişe kaydırılmış olması anlamına gelmektedir. Burada GT ifadesi Google Trend'i, GFT ifadesi ise Google Flu Trend'i ifade etmektedir. Offset işlemi tez kapsamındaki ilk çalışmada Bölüm 5.2.1'deki gibi kullanılmıştır.

Çizelge 5.6: Her hastalığın kendine ait modeli ile tahmin sonuçları

Hastalık	**R^2**	**MSE**	**Offset**
Grip	0.942	0.061	-21W, -1GFT
Listeriosis	0.271	0.559	0W, -1GT
Lyme	0.527	0.563	-2W, -2GT
Sıtma	0.137	1.008	-10W, -2GT
Boğmaca	0.517	0.616	-16W, -1GT

Beklendiği üzere, grip hastalığının tahmin edilme başarımı en yüksek çıkmış olup, Lyme ve Boğmaca hastalıklarının tahmin edilme başarımı ise grip hastalığını izlemiştir. Diğer iki hastalık için ise modellerin tahmin başarımı düşük çıkmıştır. Bu durum birkaç sebep ile açıklanabilir. Öncelikle, hastalıkların insan vücudundaki etkisi bu duruma sebep olmuş olabilir. Bazı hastalıklar vücutta çok hızlı, bazıları ise çok yavaş seyretmektedirler. Bu durum hem CDC kurumundan hem de internet ortamından gelen verinin gürültülü olmasına sebep olduğundan, bu hastalıkların tahmin edilmesini zorlaştırmıştır. Bir diğer sebep ise insanların internetteki arama alışkanlığı olabilir. İnsanlar, bazı hastalıklara yakalandıklarında kendi kendilerini tedavi etmek için internete başvururken, bazı hastalıklar için ise internete başvurmazlar. Bu ve benzeri sebeplerden ötürü, hastalıkların internet erişim ve arama verisinden yararlanmanın her hastalık için başarılı sonuç vermediğini düşünmekteyiz.

5.3.2 İkili hastalık çifti ile tahmin

Grip hastalığının diğer hastalıklara göre çok daha fazla meydana gelmesinden dolayı grip hastalığı ile alakalı hem hastane hem de internet ortamından toplanabilecek veri miktarı çok fazladır. Bu sebeple grip salgınını tahmin etmeye yönelik literatürde bir

çok çalışma gerçekleştirilmektedir. Bu bölümde, elimizde veri kümesi bulunan dört hastalığın her birinin grip hastalığı veri kümeleri ile birleştirilmiştir. Elde edilen veri kümeleri aşağıda listelenmiştir:

- Grip + Listeriosis veri kümesi
- Grip + Lyme veri kümesi
- Grip + Sıtma veri kümesi
- Grip + Boğmaca veri kümesi

Yapılan deneylerde, grip ve birleştirilen diğer hastalığın ilişkisi gözlemlenmiş ve bu iki hastalığın verisinin birleştirilmesinin hastalık salgınlarının tahmin başarımını arttırıp arttırmadığı incelenmiştir. Grip hastalığı ile diğer hastalığın semptomatik belirti yönünden bir benzerliği olduğu takdirde bu durumun en az bir hastalık tahmini için daha iyi skorlar üreteceğini hipotezi ortaya konmuştur.

Çizelge 5.7: Grip+hastalık kombinasyonları ile oluşturulan modellerin tahmin sonuçları. Her satırdaki hastalık, grip verisi ile birleştirilmiş olup kendi ve grip hastalığının sonuçlarını tahmin eder. Parantez içindeki H ifadesi o satırdaki hastalığı temsil eder.

Hastalık(H)	$\mathbf{R^2}$ (H)	$\mathbf{R^2}$ (Grip)	**MSE** (H)	**MSE** (Grip)	**Offset**
Listeriosis	0.316	0.926	0.529	0.071	0W, -1 GT
Lyme	0.561	0.936	0.520	0.069	-18W, -1 GT
Sıtma	0.159	0.914	0.976	0.092	-6W, -2 GT
Boğmaca	0.607	0.943	0.501	0.059	-21W, -1 GT

Oluşturduğumuz modelin *MSE* ve r^2 skorları Çizelge 5.7'de gösterilmiştir. Bu sonuçlardan görüldüğü üzere grip haricindeki diğer bütün hastalıkların *MSE* değerleri iyileşmiştir. Yani, bu hastalık salgınlarının tahmini için oluşturulan ikili hastalık modellerinin başarımı, Bölüm 5.3.1'de açıklanan sadece kendi veri kümesi kullanılarak ile oluşturulan model başarımdan daha iyi sonuçlar vermiştir.

Hastalıkların *MSE* değerlerindeki düşüşten başka bir önemli nokta da hastalıkların daha erkenden tahmin edilebilmesidir. Çünkü hastalık salgınları ne kadar önceden tahmin edilebilirse, bu hastalık salgınları için o kadar erkenden önlem alınabileceği anlamına gelmektedir. Bu duruma bir örnek vermek gerekirse, oluşturduğumuz ilk modelde Boğmaca hastalığı Vikipedi'nin 16 gün, Google Trend'in ise 1 hafta önceki verileri ile 0.616 *MSE* skoru ile hesaplanırken, oluşturulan ikili hastalık modelinde Vikipedi'nin 21 gün, Google Trend'in ise 1 hafta önceki verileri ile 0.501 *MSE* skoru ile tahmin edilmektedir. Bu sonuçlar baz alındığında, yeni oluşturulan deneysel model, eski modele göre hem daha başarılı hem de daha erkenden tahminler yapabilmektedir.

5.3.3 Bütün hastalıkların beraber kullanılması ile tahmin

Bölüm 5.3.2'de oluşturulan modelde görüldüğü üzere grip hastalığı veri kümesi ile diğer hastalıkların veri kümesi birleştirilip çoklu-iş öğrenme yöntemi ile hastalık salgınları için oluşturulan model başarısında artış sağlanmıştır. Bu gözlemlerimizden yola

çıkarak grip ile alakalı olabileceğini düşünerek seçtiğimiz diğer dört hastalık verisinin tümüyle ve grip verisini birleştirerek oluşturulacak çoklu-iş öğrenme modelinin başarısı incelenmiştir. Çizelge 5.8 bütün hastalıklar kullanılarak oluşturulan çoklu-iş öğrenme yöntemi model skorlarılarını göstermektedir.

Çizelge 5.8: Çoklu-iş Öğrenme(Multi-task learning) modeli tahmin sonuçları

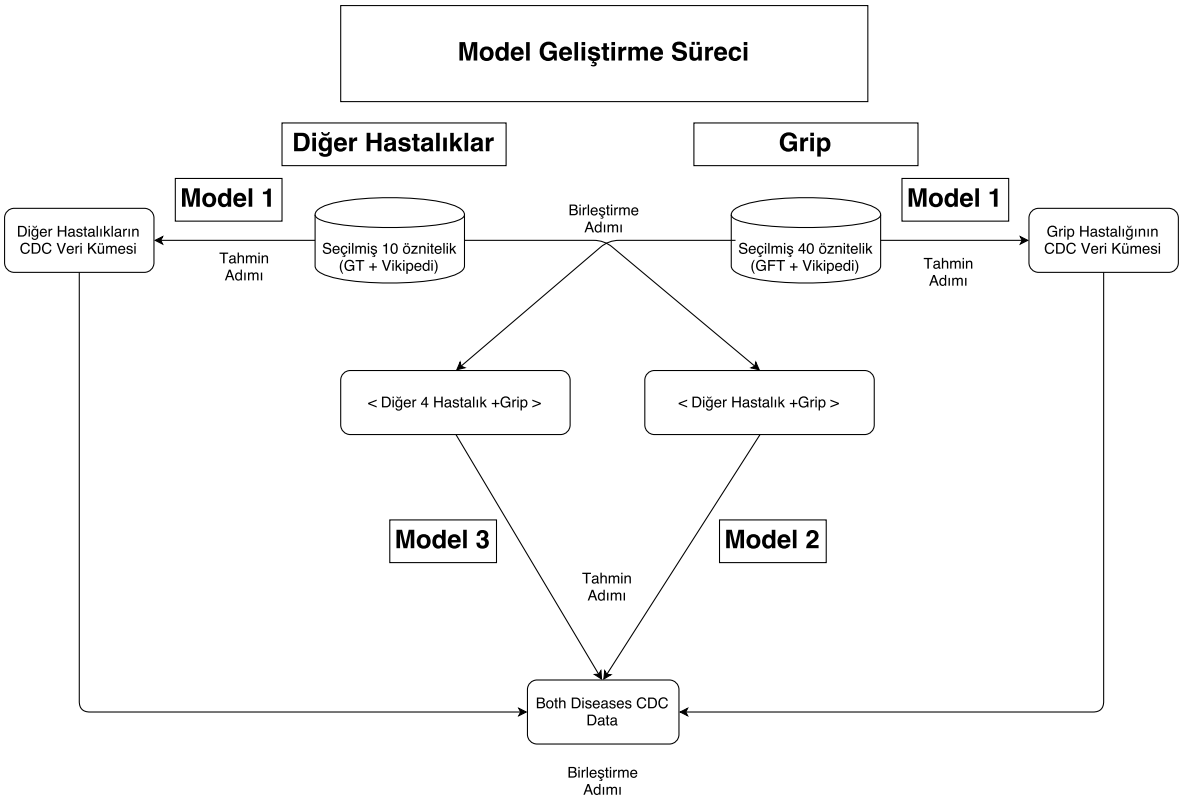
Hastalık	**R^2**	**MSE**	**En iyi Offset**
Grip	0.931	0.075	-14W, -1GT
Listeriosis	0.257	0.574	0 W, -1 GT
Lyme	0.561	0.430	0 W, -2 GT
Sıtma	0.234	0.917	-8 W, -1 GT
Boğmaca	0.647	0.485	-14W, -1 GT

Oluşturulan bu modelde Lyme, Sıtma ve Boğmaca hastalığının *MSE* değerleri daha önceden oluşturulan iki modele göre en başarılı skoru almıştır. Listeriosis hastalığı ise sadece Grip hastalığı ile beraber kullanıldığında en başarılı şekilde tahmin edilmektedir. Grip hastalığının tahmin edilme başarımında ise, oluşturulan ilk model başarımı ile karşılaştırıldığında çok küçük bir azalma olmuştur. Çoklu-iş öğrenme yöntemi veriler arasında ilişki olduğu zaman başarılı sonuçlar vermektedir. Bizim hipotezimiz ise, hastalıklar arasında semptomatik ilişki varsa bu hastalık verilerinin beraber kullanılmasıyla hastalık salgın tahmininin başarımı arttıracağı yönde olmuştur. Şekil 5.6'de görüldüğü üzere grip hastalığı haricindeki bütün hastalıkların tahmininde verilerin beraber kullanılması, hastalıkların sadece kendi internet erişim verileri ile tahmin edilmesine göre daha iyi sonuçlar vermiştir.

Tez kapsamında yapılan ikinci çalışmamızı özetleyen Şekil 5.5'da gösterilmiştir. Bu şekilde hangi modellerin hangi veri kümesi ile eğitildiği, hangi hastalıklar için hangi modellerin çalıştırıldığı gibi bilgiler ifade sergilenmiştir.

5.3.4 Tartışma

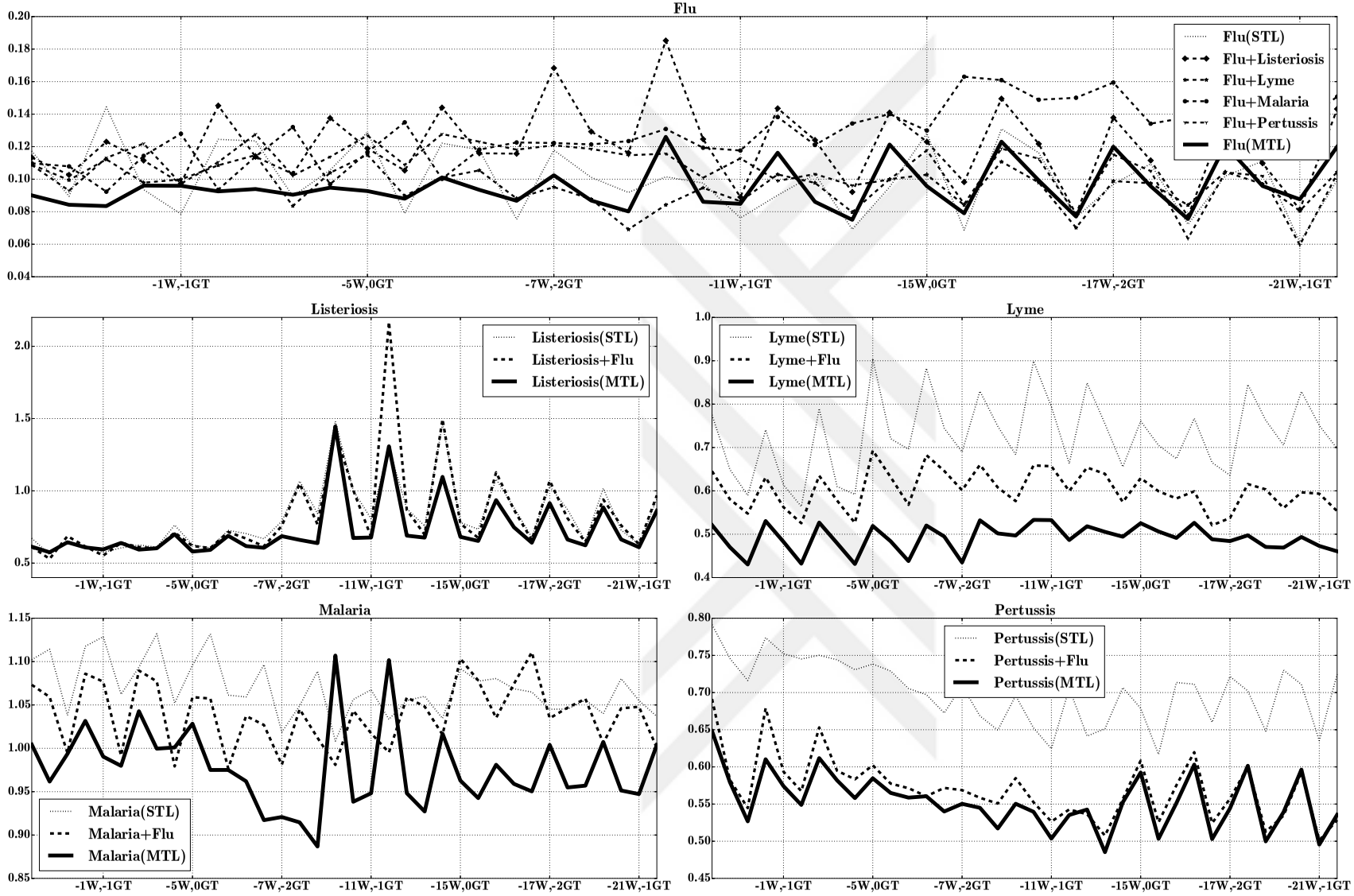
Bu çalışmada iki hipotez üzerinde çalışılmıştır. Birinci hipotez, belli hastalıklar nedeniyle hastaneye gidecek kişi sayısının, hastalıklarla alakalı internet aranma ve erişim verileri ile tahmin edilebileceği hipotezidir. Bu konu, Bölüm 2'de de örnekleri verildiği üzere başka kişiler tarafından da çalışılmıştır. İkinci hipotez ise, belli hastalıklar nedeniyle hastaneye gidecek kişi sayısının, birden fazla benzer hastalığın internet verisi ve hastane verisi birleştirilerek tahmin edilebileceği hipotezidir. Bu hipotez şu anki bilgimize göre ilk kez bu tez kapsamında ortaya konmuştur. Bu hipotezin gerçekleştirilmesi için grip hastalığı haricinde griple alakalı olabileceği düşünülen diğer hastalıkların verileri de toplanmıştır. Yapılan deneyler sonucunda hipotezimin seçilen beş hastalıktan dördü için doğru çalıştığı görülmüştür. Seçilen bu beş hastalık haricinde başka hastalıklar için de veriler toplanmış fakat istenen sonuçlar alınamamıştır. Bunun en önemli sebebi, bu hastalıkların internette aranma sıklıklarının az olması, kurumların bu hastalıklar için yeteri kadar kişi sayı bilgisi paylaşmaması gibi nedenlerden kaynaklı toplanan verinin gürültülü olmasıdır.



Şekil 5.5: 2. Çalışma model şeması

Şekil 5.6, bütün hastalıklar için, oluşturulan her üç modelinde *MSE* skorunun offset ile ilişkisini göstermektedir. Bu şekilden de açıkça görüldüğü üzere çoklu-iş öğrenme yöntemi bir hastalık haricinde diğer bütün hastalık ve offset günleri için modellerin başarımını arttırmıştır.

Her iki çalışmada da karşımıza üç farklı kısıt çıkmıştır. Birincisi, kullanılan servislerden toplanan verilerin aynı değer aralığında olmamasıdır. Google Trends, ilgili anahtar kelimenin aranma hacmini [0-100] aralığına normalize edilmiş şekilde paylaşırken, Google Flu Trends kendi algoritmasının ürettiği değeri, Vikipedi ise doğrudan makalenin erişim sayısını paylaşmaktadır. Bu durumdan dolayı yapılan veri ön işleme yöntemleri Bölüm 4'de anlatılmıştır. İkinci olarak, grip haricindeki hastalıklar için hem internet hem de hastane verisi bulabilmenin zorluğu olmuştur. Grip hastalığını tahmin etme konusunun popülerliği, hastane verilerine ulaşmanın kolaylığı ve insanların internet üzerindeki arama sıklık verisinin çok olmasından dolayı grip hastalığı için veri toplamak diğer hastalıklara göre daha kolaydır. Diğer hastalık verilerinin grip verisi kadar kaliteli şekilde toplanamıyor olması, modellerin yeterince iyi çalışmamasına neden olmuştur. Karşımıza çıkan son problem ise Vikipedi'ye ait verinin hangi lokasyona ait olduğudur. GFT ve GT servisleri paylaştığı verinin hangi lokasyona ait olduğu bilgisini verirken, Vikipedi servisinde bu bilgi mevcut değildir. Çalışmamızda İngilizce diline ait Vikipedi makaleleri kullanıldığı için bu verileri ABD'de ortaya çıkabilecek hastalık salgınlarını tahmin etmede kullanılmıştır. İngilizce dünyada başka ülkelerde de konuşulan bir dil olduğu için problem yaşayabilceğimizi düşünmemize rağmen, düşüncemize göre ABD'de teknoloji ve internetin fazla kullanılmasından dolayı Vikipedi verileri ile ABD'nin paylaştığı hastane verileri arasında önemli bir ilişki görülmüştür. Bu problem, ilk çalışmamızda da karşımıza çıkmış olup aynı şekilde yorumlanmıştır.



Şekil 5.6: Bütün deneyler için X eksininde offset değerleri, Y ekseninde *MSE* değerleri verilmiştir. Her figür, bir hastalığın 3 ayrı model ile çeşitli offset değerleri için sonuçlarını göstermektedir. Detaylı bilgi için açıklamalara bakılabilir.

6. SONUÇ

Salgın hastalıkları, özellikle grip hastalığını, tahmin etme konusu giderek dikkat çekmeye başlayan bir çalışma alanıdır. Özellikle internet ortamındaki verilerin artmasıyla beraber Vikipedi ve Google'dan alınan kullanıcı verileri ile bu tahmin işlemi daha başarılı bir şekilde gerçekleştirilmeye başlanmıştır. Bu çalışmada, Vikipedi'nin makalelerine erişim sayıları ve Google'ın çeşitli servislerinden toplanan verilerle hastalıkları tahmin eden modeller oluşturulmuştur. Toplanan bu verilerle modeller oluşturulmadan önce, verilerin birbirinden farklı yapısından dolayı bir takım ön işlemden geçirilmiştir. Veriler işlenmeye hazır hale getirildikten sonra, regresyon ve çoklu-iş öğrenme algoritmaları ile eğitilerek modeller oluşturulmuştur.

Öncelikle Vikipedi, Google Flu Trends ve Vikipedi+Google Flu Trends veri kümeleri ile modeller oluşturulup, bu modeller ile Amerika Birleşik Devletlerindeki grip seviyesi tahmin edilmeye çalışılmıştır. Literatürde, Vikipedi ve Google Flu Trends veri kümesini kullanarak çeşitli ülkelerde grip seviyesi tahmini yapan modeller oluşturulmasına rağmen bu iki veri kümesini birleştirerek tahmin işlemi yapan bir model bildiğimiz kadarıyla ilk kez bu tez çalışması kapsamında yapılmıştır. Ayrıca bu veri kümelerini beraber kullanarak oluşturulan modelin sonuçları, veri kümelerini ayrı ayrı kullanarak oluşturulan model sonuçlarından daha iyi çıkmış olup çalışmamızı ve hipotezimizi doğrulamıştır.

Tez kapsamında yapılan diğer çalışmada ise sadece grip hastalığı için değil, grip hastalığı ile ilgili olabileceğini düşündüğümüz diğer hastalıkların da görülme sıklığı tahmin edilmeye çalışılmıştır. Grip haricindeki diğer hastalıklar için Google'ın özel bir servisi olmaması nedeniyle Google Trends aracılığı ile veri kümeleri oluşturulmuştur. Literatürde, birçok hastalığın meydana gelme sayısı ile ilgili çalışmalar bulunmasına rağmen, şu anki bilgimize göre bu hastalıkların verisini beraber kullanarak tahmin etme işlemi yapılmamıştır. Modellerimizi çoklu-iş öğrenme algoritması ile eğiterek, birbirlerinin veri kümesinin ilişkisinden yararlanılması hipotezi ortaya konmuştur ve ortaya çıkan sonuçlar incelendiğinde hastalıkların verilerinin beraber kullanılması bu hipotezimizi doğrulamıştır. Ayrıca bu modellerden, hastalıkların birbiri ile ilişkisi de görülebilir olmuş olup bu da literatüre yeni bir çalışma olarak eklenmiştir.

Gelecek çalışmalarımızda, yapılan bu çalışmaları ABD haricinde başka ülkelerde ve İngilizce haricindeki başka diller ile gerçekleştirmek planlanmaktadır. Çalışmaların büyük bir kısmını kaplayan veri toplama, dönüştürme, hazır hale getirme aşamasını kısaltmak adına bu işlemleri otomatik hale getiren bir sistem tasarlamak da gelecek çalışmalarımız arasında yer almaktadır. Ayrıca tez kapsamındaki çalışmalarda kullanılmayan Twitter, Facebook gibi sosyal medya verileri ve zaman serisi modellerini gelecekteki çalışmalarımızda kullanarak daha başarılı modeller oluşturmak hedeflenmektedir.

Özetle, bu tez çalışmasının amacına ulaştığı ve internet verileri ile hastalık tahmin etme ile ilgili konuda literatüre önemli katkılar yapıldığı söylenebilir. Tez ile ilgili yapılan çalışmalar kapsamında uluslararası bir konferansta[5] yayınlanmış bir çalışma ve saygın bir dergide[6] şu an itibari ile incelemede olan çalışmalar bu durumun en güzel örneğidir.

[5]IEEE BIBE2015

[6]Journal of Biomedical and Health Informatics

KAYNAKLAR

[1] **Aitken, M.** Engaging patients through social media. *IMS Institure for Healthcare Economics*, January (2014), 1–47.

[2] **Alexa.com**. Alexa Top 500 Global Sites, 2015. Available at `http://www.alexa.com/topsites`.

[3] **Althouse, B. M., Ng, Y. Y., and Cummings, D. A. T.** Prediction of dengue incidence using search query surveillance. *PLoS Neglected Tropical Diseases 5*, 8 (2011).

[4] **Carneiro, H. A., and Mylonakis, E.** Google Trends: A Web-Based Tool for Real-Time Surveillance of Disease Outbreaks. *Clinical Infectious Diseases 49*, 10 (2009), 1557–1564.

[5] **Caruana, R.** Multitask Learning. *Machine Learning 28*, 1 (1997), 41–75.

[6] **Chan, E. H., Sahai, V., Conrad, C., and Brownstein, J. S.** Using web search query data to monitor dengue epidemics: A new model for neglected tropical disease surveillance. *PLoS Neglected Tropical Diseases 5*, 5 (2011).

[7] **Cho, S., Sohn, C. H., Jo, M. W., Shin, S. Y., Lee, J. H., Ryoo, S. M., Kim, W. Y., and Seo, D. W.** Correlation between national influenza surveillance data and Google Trends in South Korea. *PLoS ONE 8*, 12 (2013).

[8] **Choi, H., and Varian, H.** Predicting the Present with Google Trends. *Economic Record 88*, SUPPL.1 (2012), 2–9.

[9] **Clauson, K. a., Polen, H. H., Boulos, M. N. K., and Dzenowagis, J. H.** Accuracy and completeness of drug information in Wikipedia. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium 99*, October (2008), 912.

[10] **Desai, R., Hall, A. J., Lopman, B. A., Shimshoni, Y., Rennick, M., Efron, N., Matias, Y., Patel, M. M., and Parashar, U. D.** Norovirus disease surveillance using google internet query share data. *Clinical Infectious Diseases 55*, 8 (2012).

[11] **Dugas, A. F., Jalalpour, M., Gel, Y., Levin, S., Torcaso, F., Igusa, T., and Rothman, R. E.** Influenza Forecasting with Google Flu Trends. *PLoS ONE 8*, 2 (2013).

[12] **Evgeniou, T., and Pontil, M.** Regularized multi-task learning. *International Conference on Knowledge Discovery and Data Mining* (2004), 109.

[13] **Fantazzini, D., and Toktamysova, Z.** Forecasting German car sales using Google data and multivariate models. *International Journal of Production Economics 170, Part* (2015), 97–135.

[14] **Generous, N., Fairchild, G., Deshpande, A., Del Valle, S. Y., and Priedhorsky, R.** Global Disease Monitoring and Forecasting with Wikipedia. *PLoS Computational Biology 10*, 11 (2014).

[15] **Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L.** Detecting influenza epidemics using search engine query data. *Nature 457*, 7232 (2009), 1012–4.

[16] **Hickmann, K. S., Fairchild, G., Priedhorsky, R., Generous, N., Hyman, J. M., Deshpande, A., and Del Valle, S. Y.** Forecasting the 2013–2014 Influenza Season Using Wikipedia. *PLOS Computational Biology 11* (2015), e1004239.

[17] **Hunter, J. D.** Matplotlib: A 2D graphics environment. *Computing in Science and Engineering 9*, 3 (2007), 99–104.

[18] **Jones, E., Oliphant, T., Peterson, P., et al.** SciPy: Open source scientific tools for Python, 2001–. [Online; accessed 2016-08-03].

[19] **Kman, N. E., and Bachmann, D. J.** Biosurveillance a review and update. *Advances in preventive medicine 2012* (2012), 301408.

[20] **Lazer, D., Kennedy, R., King, G., and Vespignani, A.** The Parable of Google Flu: Traps in Big Data Analysis. *Science 343*, 6167 (2014), 1203–1205.

[21] **Leithner, A., Maurer-Ertl, W., Glehr, M., Friesenbichler, J., Leithner, K., and Windhager, R.** Wikipedia and osteosarcoma: a trustworthy patients' information? *Journal of the American Medical Informatics Association : JAMIA 17*, 4 (2010), 373–374.

[22] **Li, X., Ma, J., Wang, S., and Zhang, X.** How does Google search affect trader positions and crude oil prices? *Economic Modelling 49*, March 2013 (2015), 162–171.

[23] **McIver, D. J., and Brownstein, J. S.** Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time. *PLoS Computational Biology 10*, 4 (2014).

[24] **McKinney, W.** pandas: a Foundational Python Library for Data Analysis and Statistics. *Python for High Performance and Scientific Computing* (2011), 1–9.

[25] **Mestyán, M., Yasseri, T., and Kertész, J.** Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data. *PLoS ONE 8*, 8 (2013).

[26] **Moat, H. S., Curme, C., Avakian, A., Kenett, D. Y., Stanley, H. E., and Preis, T.** Quantifying Wikipedia Usage Patterns Before Stock Market Moves, 2013.

[27] **Mohebbi, M., Vanderkam, D., Kodysh, J., Schonberger, R., Hyunyoung, C., and Kumar, S.** Google Correlate Whitepaper. *Google* (2011), 1–6.

[28] **Molinari, N. A. M., Ortega-Sanchez, I. R., Messonnier, M. L., Thompson, W. W., Wortley, P. M., Weintraub, E., and Bridges, C. B.** The annual impact of seasonal influenza in the US: Measuring disease burden and costs. *Vaccine 25*, 27 (2007), 5086–5096.

[29] **Ng, A. Y.** Feature selection, L1 vs. L2 regularization, and rotational invariance. *Twenty-first international conference on Machine learning - ICML '04* (2004), 78.

[30] **Ocampo, A. J., Chunara, R., and Brownstein, J. S.** Using search queries for malaria surveillance, Thailand. *Malaria journal 12* (2013), 390.

[31] **Olson, D. R., Konty, K. J., Paladini, M., Viboud, C., and Simonsen, L.** Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographic Scales. *PLoS Computational Biology 9*, 10 (2013).

[32] **Ortiz, J. R., Zhou, H., Shay, D. K., Neuzil, K. M., Fowlkes, A. L., and Goss, C. H.** Monitoring Influenza activity in the United States: A comparison of traditional surveillance systems with Google Flu Trends. *PLoS ONE 6*, 4 (2011).

[33] **Pedregosa, F., and Varoquaux, G.** Scikit-learn: Machine learning in Python. ... *of Machine Learning ... 12* (2011), 2825–2830.

[34] **Rajagopalan, M. S., Khanna, V. K., Leiter, Y., Stott, M., Showalter, T. N., Dicker, A. P., and Lawrence, Y. R.** Patient-oriented cancer information on the internet: a comparison of wikipedia and a professionally maintained database. *Journal of oncology practice / American Society of Clinical Oncology 7*, 5 (2011), 319–23.

[35] **Ram, S., Zhang, W., Williams, M., and Pengetnze, Y.** Predicting Asthma-Related Emergency Department Visits Using Big Data. *IEEE Journal of Biomedical and Health Informatics 19*, 4 (2015), 1216–1223.

[36] **Romera-Paredes, B., Argyriou, A., Pontil, M., Berthouze, N., and Pontil, M.** Exploiting Unrelated Tasks in Multi-Task Learning. *Proc. 15th Int. Conf. Artificial Intell. and Stat. 22* (2012), 951–959.

[37] **Seifter, A., Schwarzwalder, A., Geis, K., and Aucott, J.** The utility of "Google Trends" for epidemiological research: Lyme disease as an example. *Geospatial Health 4*, 2 (2010), 135–137.

[38] **Shaman, J., Pitzer, V. E., Viboud, C., Grenfell, B. T., and Lipsitch, M.** Absolute humidity and the seasonal onset of influenza in the continental United States. *PLoS Biology 8*, 2 (2010).

[39] **Tausczik, Y., Faasse, K., Pennebaker, J. W., and Petrie, K. J.** Public Anxiety and Information Seeking Following the H1N1 Outbreak: Blogs, Newspaper Articles, and Wikipedia Visits. *Health Communication 27*, 2 (2012), 179–185.

[40] **Tibshirani, R.** Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society. Series B: Statistical Methodology 73*, 3 (2011), 273–282.

[41] **Trends, G.**, 2016. Available at `https://www.google.com/trends/` [Online; accessed 3-January-2016].

[42] **Trends, G. F.**, 2015. Available at `https://www.google.org/flutrends/about/data/flu/us/data.txt` [Online; accessed 10-April-2015].

[43] **Urllib**. Urllib library, 2016. Available at `https://docs.python.org/2/library/urllib.html`.

[44] **Weisstein, Eric W.** Frobenius Norm. *MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/FrobeniusNorm.html* (2011).

[45] **Who**. Who_factsheets, 2011. Available at `www.who.int/mediacenter/factsheets/fs211/en/`.

[46] **Wikipedia**. Centers for disease control and prevention — wikipedia, the free encyclopedia, 2016. Available at `https://en.wikipedia.org/w/index.php?title=Centers_for_Disease_Control_and_Prevention&oldid=732567377`.

[47] **Wilson, K., and Brownstein, J. S.** Early detection of disease outbreaks using the Internet. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne 180*, 8 (2009), 829–831.

[48] **Yang, S., Santillana, M., and Kou, S. C.** Accurate estimation of influenza epidemics using Google search data via ARGO. *Proceedings of the National Academy of Sciences of the United States of America 112*, 47 (2015), 14473–14478.

[49] **Zou, H., and Hastie, T.** Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology 67*, 2 (2005), 301–320.

ÖZGEÇMİŞ

Ad-Soyad : Batuhan Bardak

Uyruğu : T.C.

Doğum Tarihi ve Yeri : 12.07.1991 - Ankara

E-posta : b.bardak@etu.edu.tr

ÖĞRENİM DURUMU:

- **Lisans** : 2014, TOBB ETU, Bilgisayar Mühendisliği

MESLEKİ DENEYİM VE ÖDÜLLER:

Yıl	Yer	Görev
2014-2016	TOBB ETU	Burslu Yüksek Lisans Öğrencisi
2014-2014	TAI - TUSAŞ	Stajyer
2013-2013	Valeo Schalter und Sensoren GmbH	Stajyer
2012-2012	Sistemim IT	Stajyer

YABANCI DİL: İNGİLİZCE

TEZDEN TÜRETİLEN YAYINLAR, SUNUMLAR VE PATENTLER:

- B. Bardak and M. Tan, "Disease outbreak prediction by data integration and multi-task learning" (Under review)
- B. Bardak and M. Tan, "Prediction of influenza outbreaks by integrating Wikipedia article access logs and Google flu trend data," Bioinformatics and Bioengineering (BIBE), 2015 IEEE 15th International Conference on, Belgrade, 2015, pp. 1-6. doi: 10.1109/BIBE.2015.7367640